Institute of Metabolic Science

MRC | Epidemiology Unit

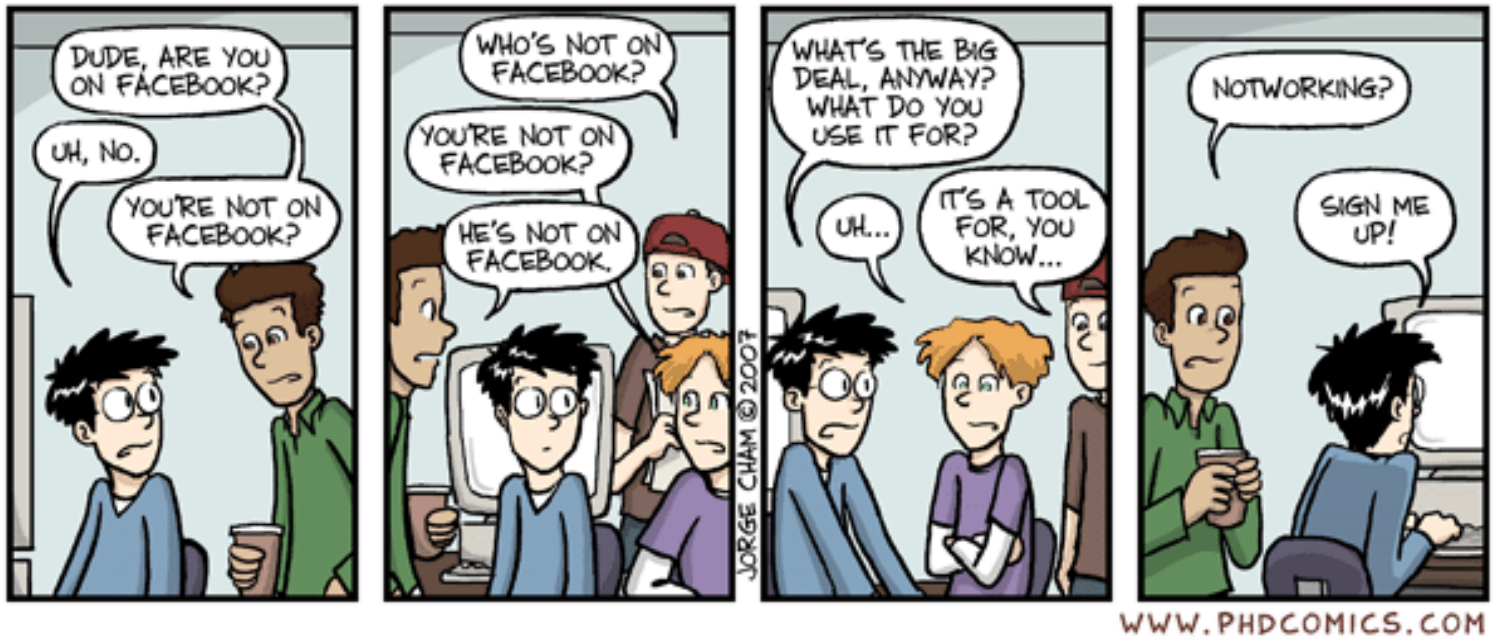
Genome-wide Association Studies (GWASs)

July 7, 2009, Agrocampus-Ouest, Rennes, France

Jing Hua Zhao

2009

A self-introduction



A sketch

- We aim to give a brief overview of genetic analysis of complex traits and diseases in humans in the context of large volume of data.
- Our focus is on practical issues.
- We provide specific examples of genetic association study.
- We are not limited to R.
- We hope this will serve as a forum for a range of issues and a contact point for future researches.

Structure

We split our time into four parts each about 45min:

1. Overview
2. Association testing
3. Meta-analysis
4. Other topics

You may find materials from useR! 2008 tutorial relevant.

Institute of Metabolic Science

MRC | Epidemiology Unit

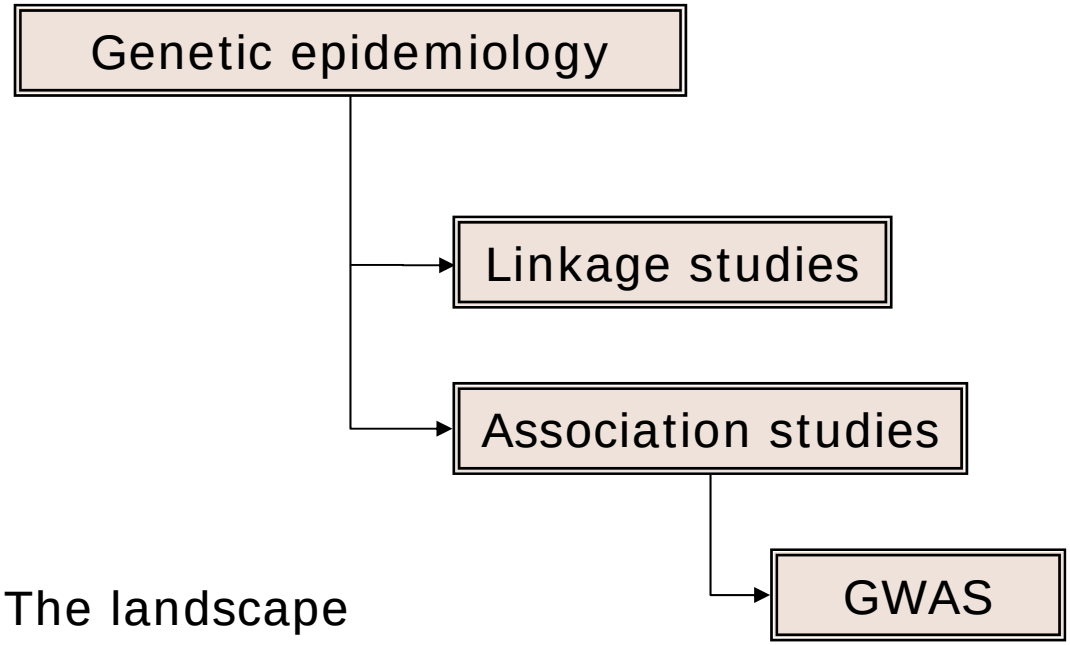
Genome-wide Association Studies

Overview

Jing Hua Zhao

Topics

- Organization



- The landscape
- Study design
- Analytical tools

Genetic epidemiology

- It is the study of the role of genetic factors in determining health and disease in families and in populations, and the interplay of such genetic factors with environmental factors, or “a science which deals with the aetiology, distribution, and control of diseases in groups of relatives and with inherited causes of disease in populations” (http://en.wikipedia.org).
- It customarily includes study of familial aggregation, segregation, linkage and association. It is closely associated with the development of statistical methods for human genetics which deals with these four questions. The last two questions can only be answered if appropriate genetic markers available (Elston & Ann Spence. *Stat Med* 2006;25:3049-80).

Some terminology

- Genes, Chromosome, markers
- Alleles, genotypes, haplotypes
- Phenotypes, mode of inheritance, penetrance
- Mendelian laws of inheritance, Hardy-Weinberg equilibrium, linkage disequilibrium
- Association tests for single or multiple SNPs
- Population stratification
- Multiple testing
- Gene-environment interaction

Linkage studies

- It is the study of cosegregation between genetic markers and putative disease loci, and has been very successful in localizing rare, Mendelian disorders but since has difficulty for traits which do not strictly follow Mendelian mode of inheritance, considerable linkage heterogeneity and it has limited resolution.
- It typically involves parametric (model-based) and nonparametric (model-free) methods, the latter most commonly refers to allele-sharing methods.
- The underlying concepts are nevertheless very important. It can still be useful in providing candidates for fine-mapping and association studies.
- With availability of whole genome data, it is possible to infer relationship or correlation between any individuals in a population.

Association studies

- They focus on association between particular allele and trait; it is only feasible with availability of dense markers.
- It has traditionally applied to both relatives in families and population sample. For the latter there has been serious concern over spurious association due to difference in allele frequencies between hidden sub-populations in a sample.
- A range of considerations has been made (Balding. *Nat Rev Genet* 2006;7:781-91) but the availability of whole genome data again refresh views including statistical examination of population substructure.

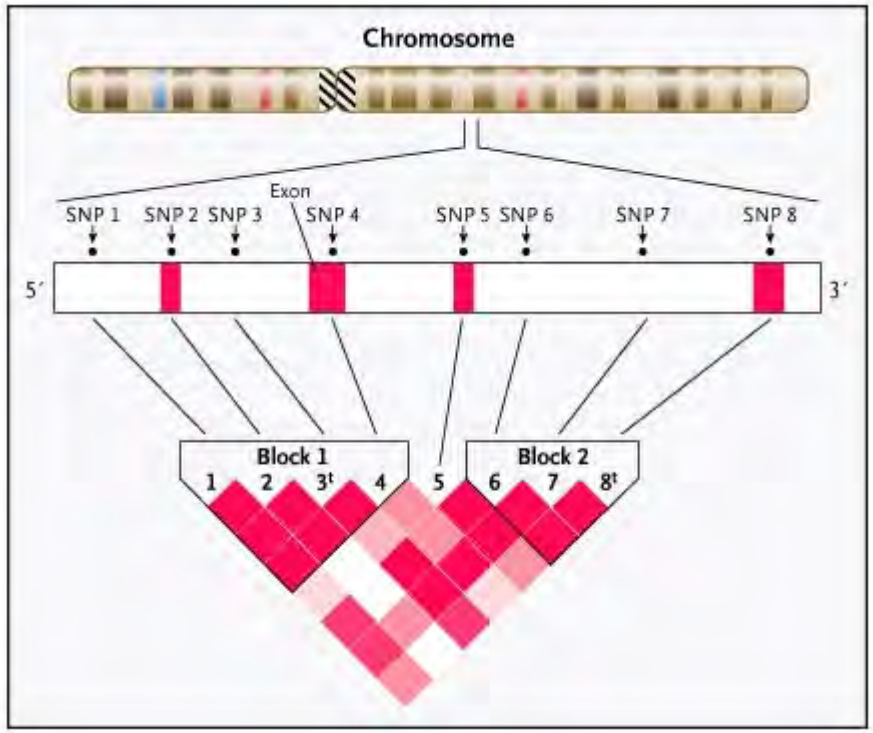
GWAS and assumptions

- Any study of genetic variation across the entire human genome designed to identify genetic association with observable traits or the presence or absence of a disease, usually referring to studies with genetic marker density of 100,000 or more to represent a large proportion of variation in the human genome (Pearson & Manolio. *JAMA* 2008;299:1335-44), or simply … look for associations between DNA sequence variants and phenotypes of interest (Donnelly. *Nature 2008;* 456:728-31).
- The logic is associated with the so-called common disease common variant hypothesis (CD-CV). Common polymorphisms (MAF>1%) might contribute to susceptibility to common diseases, so that GWAS of common variants might be used to map loci contributing to common diseases. It therefore helps to catalog millions of common variants in the human population, massive genotypes to large number of individuals, and appropriate analytical framework (Altshuler et al. *Science* 2008; 322:881-888).

The landscape of GWASs

- A catalog of published GWASs is maintained by Office of Population Genomics at the National Human Genome Research Institute (NHGRI) and available from http://www.genome.gov/26525384
- As of 03/03/09, the table includes 273 publications and 1213 SNPs. For instance, for body mass index, it includes all the major publications from GWASs, namely, Thorleifsson et al. *Nat Genet* 2009;41:18-24, Willer et al. *Nat Genet* 2009; 41:25-34; Loos et al. *Nat Genet* 2008;40:768-75, Fox et al. *BMC Med Genet* 2007;8:S18, Frayling et al. *Science* 2007;316:889-94.
- Furthermore, there were Benzinou et al. *Nat Genet* 2008;40:943-5, Meyre et al. *Nat Genet* 2009;41:157-9.

Context of GWASs



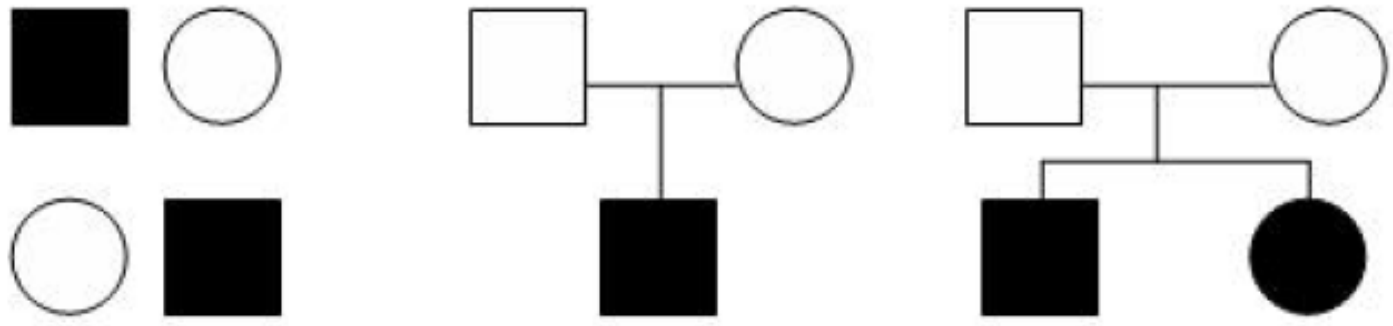
- The population under study must be characterized to allow the selection of patients likely to share a genetic cause of disease
- Thousands of cases and controls may be needed if a study is to have sufficient statistical power to identify the alleles of interest
- … it creates bioinformatics challenges and raises questions about how to identify true positive signals

Christensen & Murray. *New Eng J Med* 2007;356:1094-7

International collaborative projects

- The HapMap project (http://www.hapmap.org) is a study of 270 people from the Yoruba in Nigeria (30 trios), Japanese (45 unrelated individuals), Han Chinese (45 unrelated individuals) and CEPH (30 trios).
- The 1000 genome project (http://www.1000genomes.org) aims to sequence at least one thousand anonymous participants.
- The database of genotypes and phenotypes (dbGaP) (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.
- The genetic analysis workshops (GAWs) (http://www.gaworkshop.org/) are a collaborative effort among genetic epidemiologists to evaluate and compare statistical genetic methods. For each GAW, topics are chosen that are relevant to current analytical problems through simulated or real data.

Study designs



- Three common genetic association designs involving unrelated individuals (left), nuclear families with affected singletons (middle) and affected sib-pairs (right). Males and females are denoted by squares and circles with affected individuals filled with black colors and unaffected individuals being empty
- Risch & Merikangas. Science 1996;273:1516-7, Zhao. *J Stat Soft* 2007;23(8):1-18

Sample sizes required for association detection using population data

		K			
γ	p	1%	5%	10%	20%
4.0	0.01	46638	8951	4240	1885
	0.10	8173	1569	743	331
	0.50	10881	2089	990	440
	0.80	31444	6035	2859	1271
2.0	0.01	403594	77458	36691	16307
	0.10	52660	10107	4788	2128
	0.50	35252	6766	3205	1425
	0.80	79317	15223	7211	3205
1.5	0.01	1598430	306770	145312	64583
	0.10	191926	36835	17448	7755
	0.50	97922	18793	8902	3957
	0.80	191926	36835	17448	7755

Power of linkage versus association

		Linkage			Association					
γ	p	Y	N_L	P_A	H_1	N_{tdt}	H_2	$N_{asp/tdt}$	λ_o	λ_s
4.00	0.01	0.520	6400	0.800	0.048	1098	0.112	235	1.08	1.09
	0.10	0.597	277	0.800	0.346	151	0.537	48	1.48	1.54
	0.50	0.576	445	0.800	0.500	104	0.424	62	1.36	1.39
	0.80	0.529	3023	0.800	0.235	223	0.163	162	1.12	1.13
2.00	0.01	0.502	445839	0.667	0.029	5824	0.043	1970	1.01	1.01
	0.10	0.518	8085	0.667	0.245	696	0.323	265	1.07	1.08
	0.50	0.526	3752	0.667	0.500	340	0.474	180	1.11	1.11
	0.80	0.512	17904	0.667	0.267	640	0.217	394	1.05	1.05
1.50	0.01	0.501	6942837	0.600	0.025	19321	0.031	7777	1.00	1.00
	0.10	0.505	101898	0.600	0.214	2219	0.253	941	1.02	1.02
	0.50	0.510	27041	0.600	0.500	950	0.490	485	1.04	1.04
	0.80	0.505	101898	0.600	0.286	1663	0.253	941	1.02	1.02

Power calculation for multiple regression

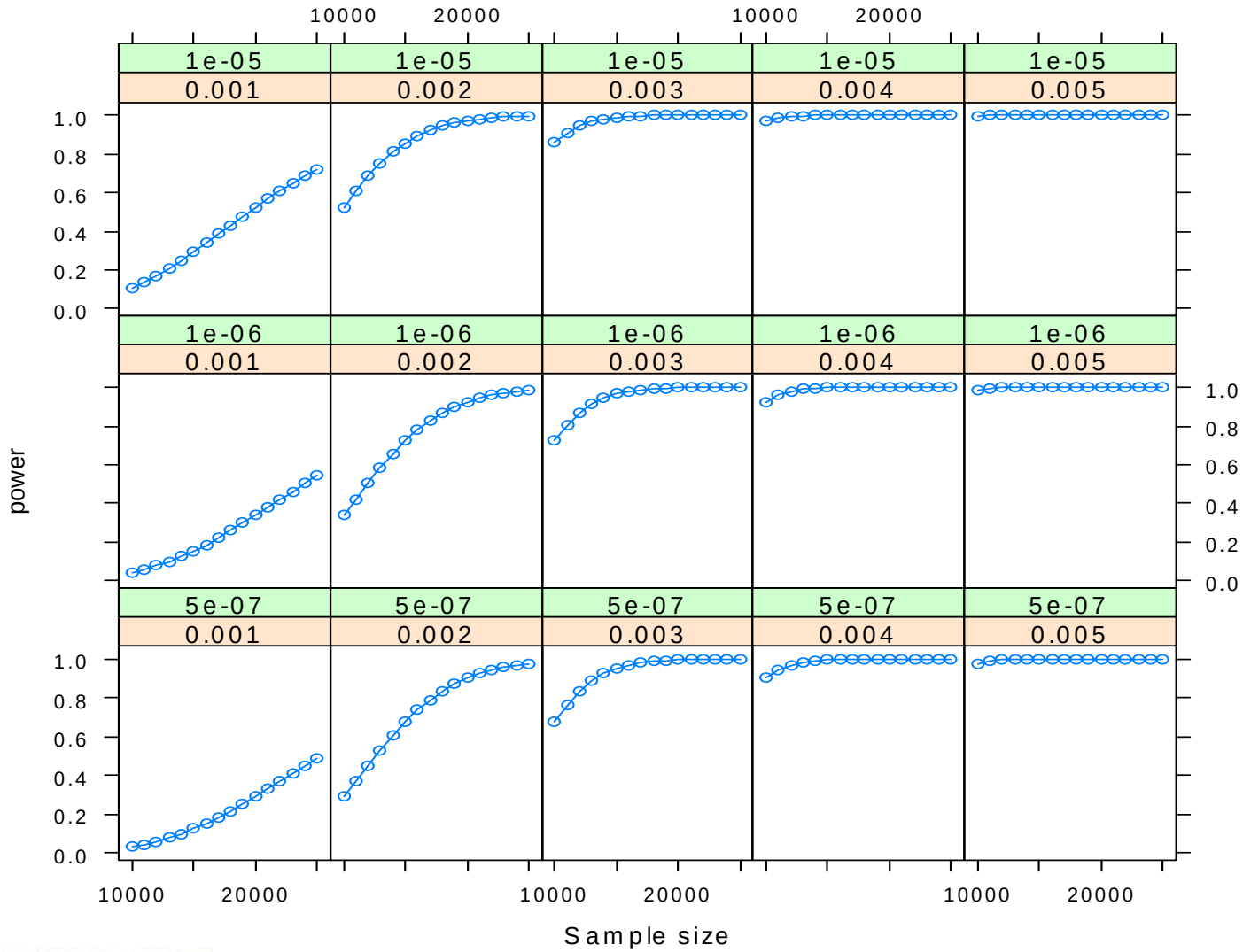
```
proc power;
    ods output output=op;
    multreg
        model = fixed
        alpha = 0.00001 0.000001 0.0000005
        nfullpred = 1
        ntestpred = 1
        rsqfull = 0.001 to 0.005 by 0.001
        rsqdiff = 0.001 to 0.005 by 0.001
        ntotal = 10000 to 25000 by 1000
        power = .;
run;
```

- This is according to a linear model with given proportions of variance explained, significant levels and sample sizes.

Power by %variance explained (R^2)

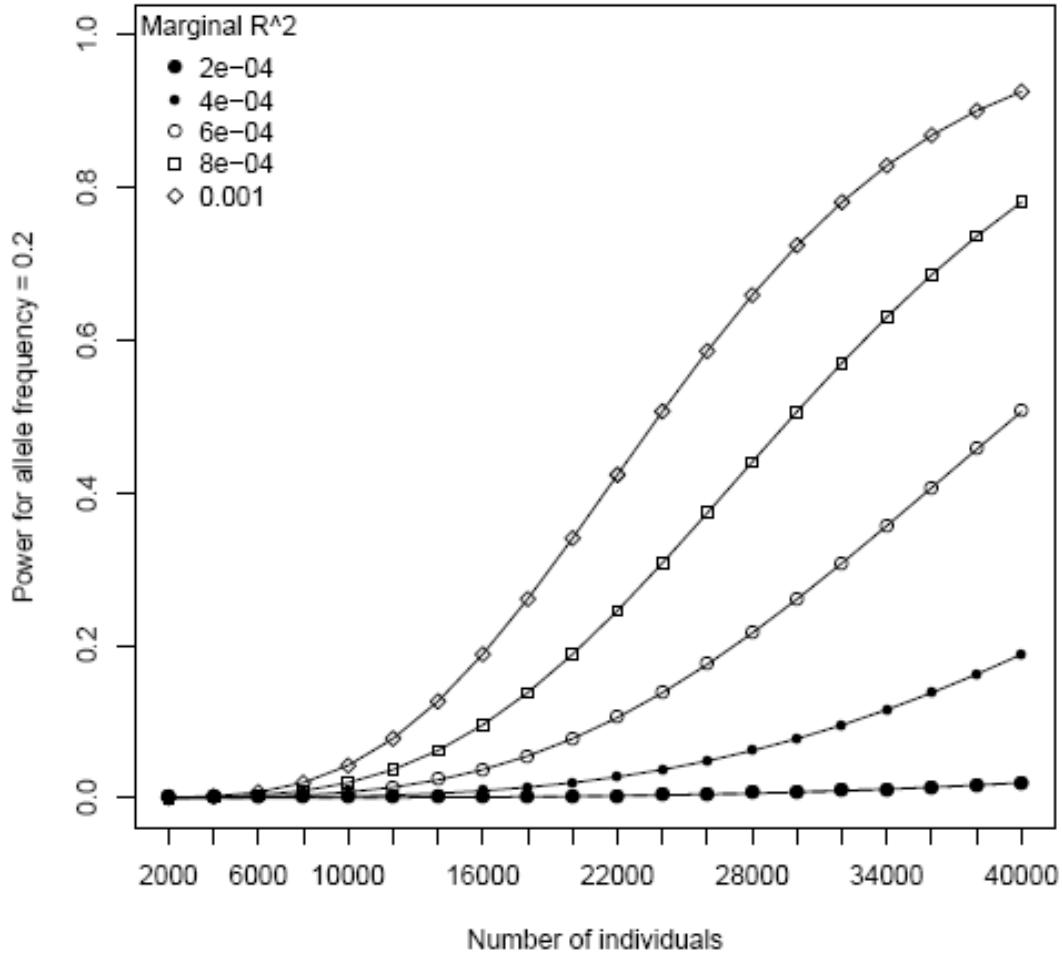
	R^2				
Sample size	0.1	0.2	0.3	0.4	0.5
$\alpha = 10^{-5}$					
10000	0.10	0.52	0.86	0.97	1
15000	0.29	0.86	0.99	1.00	1
20000	0.52	0.97	1.00	1.00	1
25000	0.72	1.00	1.00	1.00	1
$\alpha = 5x10^{-7}$					
10000	0.031	0.29	0.67	0.9	0.98
15000	0.124	0.67	0.95	1.0	1.00
20000	0.290	0.90	1.00	1.0	1.00
25000	0.489	0.98	1.00	1.0	1.00

Power by α =1e-5,1e-6,5e-7



Gene-Environment interaction (GEI)

- This was prepared in response to a call for establishing large cohort.
- Assume a quantitative trait and an environment factor both with N(0,1) and an additive model for genetic effect, the power for minor allele frequency=0.2, significant level 1e-6 is shown.



Two-stage GEI

- A case-only design is used as the first stage.
- This is to be followed by a second stage involving both cases and controls.

Kass & Gold. *Handbook of Epidemiology* 2004; I.7

Murcray et al. *Am J Epidemiol* 2008; 169:219-26

Li & Conti. *Am J Epidemiol* 2008; 169:497-504

Software for data analysis

- LINKAGE, GENEHUNTER, Merlin, PAP, SAGE, SOLAR
- ETDT, FBAT, QTDT, UNPHASED, SAS/Genetics
- R (genetics, gap, haplo.stats, haplin, kinship)
- See reviews on *Human Genomics*
- A comprehensive list is available from http://linkage.rockefeller.edu

Software for GWAS

- HaploView
- PLINK, SNPGWA
- IMPUTE, MACH, BinBam
- EIGENSTRAT
- METAL
- SAS, Stata, R (snpMatrix, GenABEL, SNPassoc)

Connections

- Occasionally, these software will be referenced.
- Analyses with specialized programs such as IMPUTE/SNPTEST and PLINK are illustrated in the useR!2008 tutorial.
- It appears that data could be used within snpMatrix, e.g., narac <- read.plink("narac.bed","narac.bim","narac.fam")

SAS

- Procedures in SAS/BASE and other modules provide graphics, database support and internet connectivity.
- SAS/STAT provides standard procedures including linear and logistic regressions or generalized linear (nonlinear, mixed) model as well as covariance and linear structure model (CALIS), MULTTEST.
- SAS/Genetics includes procedures for summarizing marker data (ALLELE), inferring and tagging haplotypes (HAPLOTYPE and HTSNP), association testing in population-based (CASECONTROL) and family-based (FAMILY) samples.

Stata

- It is a general-purpose, modern and easy to use statistical system.
- There is a good implementation for meta-analysis (metan, etc).
- There are a set of functions for instrumental variable regressions.
- Available implementations for genetic data includes summary statistics, test of Hardy-Weinberg equilibrium, haplotype estimation, tagging and association analysis.

stpower logrank

- **The effect of interferon gamma-1b on survival in patients with idiopathic pulmonary fibrosis: A multinational randomized placebo-controlled trial**
- The study was designed to provide 90 percent power to detect a treatment effect equivalent to a 50 percent reduction (i.e., from 24 percent to 12 percent) in 3-year mortality using a log-rank statistic with a 0.025 one-sided alpha. Sample size calculations indicated that approximately 600 patients would be required to achieve the targeted number of events within the planned duration of the trial.". It would be helpful to elaborate how the percentages were derived. From Stata's *stpower logrank* this would be equivalent to a withdrawal probability of 15 percent and Schoenfeld's formula.

References

- Armitage P, Colton T. Encyclopedia of Biostatistics, Second Edition, Wiley 2005
- Balding DJ, Bishop M, Cannings C. Handbook of Statistical Genetics, Third Edition, Wiley 2007
- Elston RC, Johnson W. Basic Biostatistics for Geneticists and Epidemiologists: A Practical Approach. Wiley 2008
- Haines JL, Pericak-Vance M. Genetic Analysis of Complex Diseases, Second Edition. Wiley 2006
- Rao DC, Gu CC (Eds). Genetic Dissection of Complex Traits, Volume 60, Second Edition (Advances in Genetics). Academic Press 2008
- Thomas DC. Statistical Methods for Genetic Epidemiology. Oxford University Press 2004

A summary

- We have to restricted our focus and leave out a lot of details to cover a rapid moving field with limited time.
- It seems that the practice of study designs and data analysis cannot be changed in a short run, but we have already seen steady increase in use of R.

Case study: GWAS of obesity-related traits

- Background
- Study design
- Statistical analysis
- On-going research

EPIC study

The European Prospective Investigation into Cancer and Nutrition (EPIC) is coordinated by Dr Elio Riboli, Head of the Division of Epidemiology, Public Health and Primary Care at the Imperial College London.

EPIC was designed to investigate the relationships between diet, nutritional status, lifestyle and environmental factors and the incidence of cancer and other chronic diseases. EPIC is the largest study of diet and health ever undertaken, having recruited over half a million (520,000) people in ten European countries: Denmark, France, Germany, Greece, Italy, The Netherlands, Norway, Spain, Sweden and the United Kingdom.

EPIC-Norfolk study

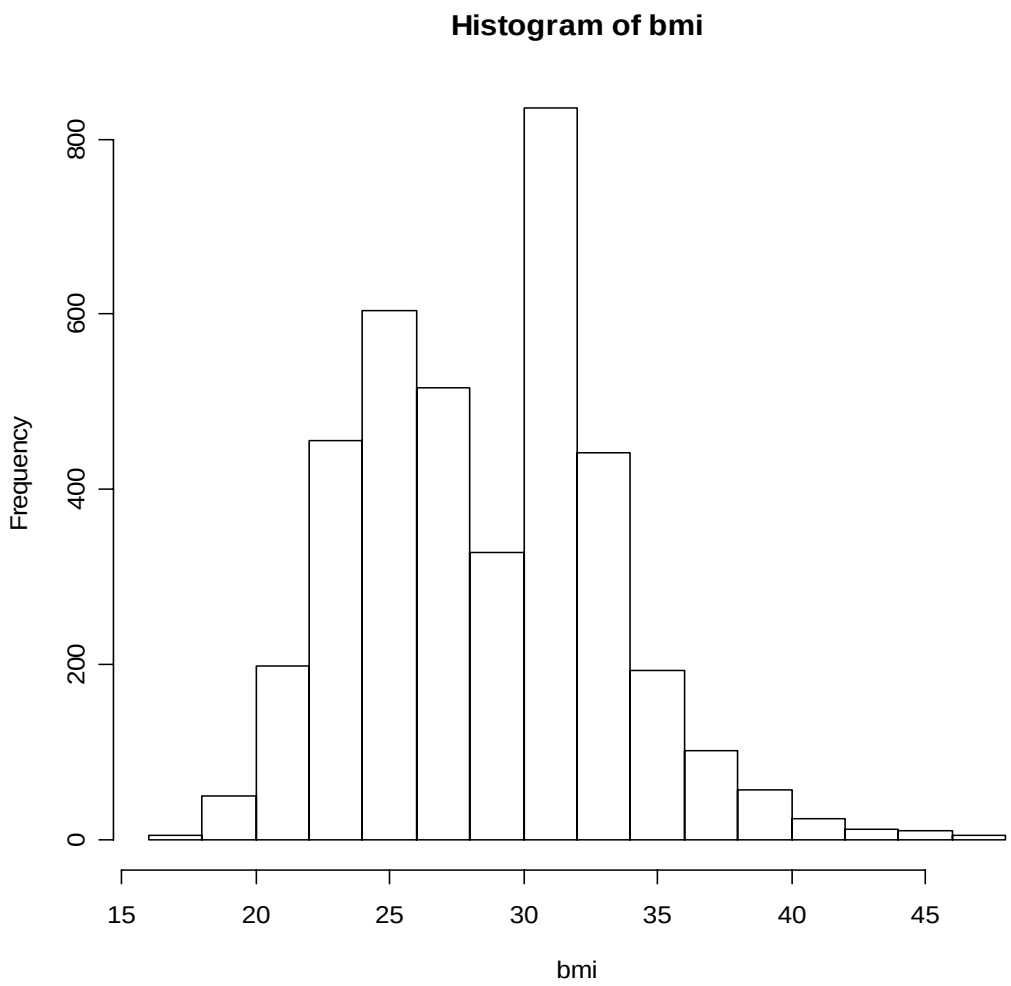
EPIC-Norfolk participants are men and women (based on over 30,000 people) who were aged between 45 and 74 when they joined the study, who lived in Norwich and the surrounding towns and rural areas. They have been contributing information about their diet, lifestyle and health through questionnaires, and through health checks carried out by EPIC nurses.

Case-cohort design for EPIC-Norfolk study

- It originally followed case-control design (e.g., WTCCC with seven cases and common controls) with 3425 cases and 3400 controls.
  - It is potentially more powerful.
  - Controls are selected.
- It has then been changed into case-cohort design, in which cases are defined to be individuals whose BMI above 30 and controls are a random sample (subcohort) of the EPIC-Norfolk cohort which includes obese individuals.
  - The subcohort is representative of the whole population and allows for a range of traits to be examined.
  - The analysis is potentially more involved but established.

Case-cohort design

- The distribution of body mass index (BMI) is the case-cohort design of the EPIC-Norfolk study of obesity is a combination of the sub-cohort sample and case sample which is truncated from the whole cohort at BMI=30
- Zhao. *J Stat Soft* 2007;23(8):1-18



Power/sample size

- It started with assessment of how the power is compromised relative to the original case-control design.
- This was followed by power/sample size calculation using methods established by Cai and Zeng (2004) as implemented in an R function, noting a number of assumptions.
- More practically, it was also envisaged that a proper representative sample of a total of 25,000 individuals would be 10%; the subcohort is then approximately 2,500.
- The total sample was split between two stages.

GeneChips

- Affymetrix 500K
  - Data were available for 3850 individuals
- Illumina 317K
  - It came at a later time。
  - Data quality appears to be poor?
- The focus has therefore been Affy500K, but with a possible comeback.

Analysis

- An incremental approach was adopted since the storage and computing power were somewhat uncertain.
- This was predated with controls from the breast cancer study, involving about 400 individuals with Perlegen 250K GeneChips.
- QC including call rates and HWE was feasible with SAS/Genetics (~30GB) which provides a good estimate of the storage for all individuals (~380GB).
- The Linux platform seemed to be favourable.

EPIC400 analysis

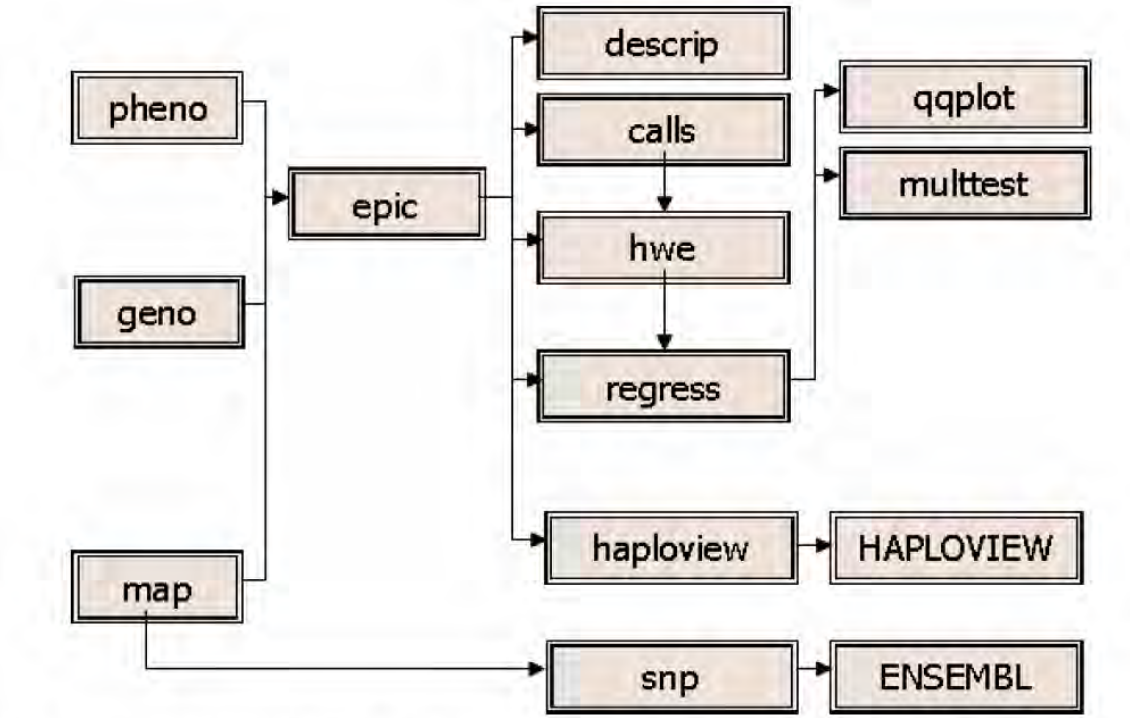


Fig. 1. A flowchart of the EPIC 400 analysis, with modules in brackets. Genotypes (geno) and phenotypes (pheno) are merged (epic) for descriptive statistics (descrip) call rates (calls), HWE (hwe), regression (regress) with adjustment for multiple testing (multtest) and comparison with theoretical distribution (qqplot). The raw data together with map information (map) can also be reformatted (haploview) into HAPLOVIEW input files so that specific region in the genome can be visualized, with annotation information from ENSEMBL according to SNPs (snp).

The analysis for GWAS

- QC including visualisation of clustering, outliers, was largely done by colleagues at Sanger (as for WTCCC)
- The overall strategy was data partition, i.e., by chromosome and further by region (30) in each chromosome, largely on a long, skinny data format
- A major advantage is that the analysis can be resumed whenever the system experiences problems
- We stuck to SAS to allow for reliability and flexibility with or without SAS/Genetics, for BMI/obesity as continuous and binary outcomes are readily tackled with REG/LOGISTIC procedures – most outputs are available from the output delivery system (ODS)
- The picture was eventually changed with a revised coding algorithm and the use of imputed data

Additional analysis

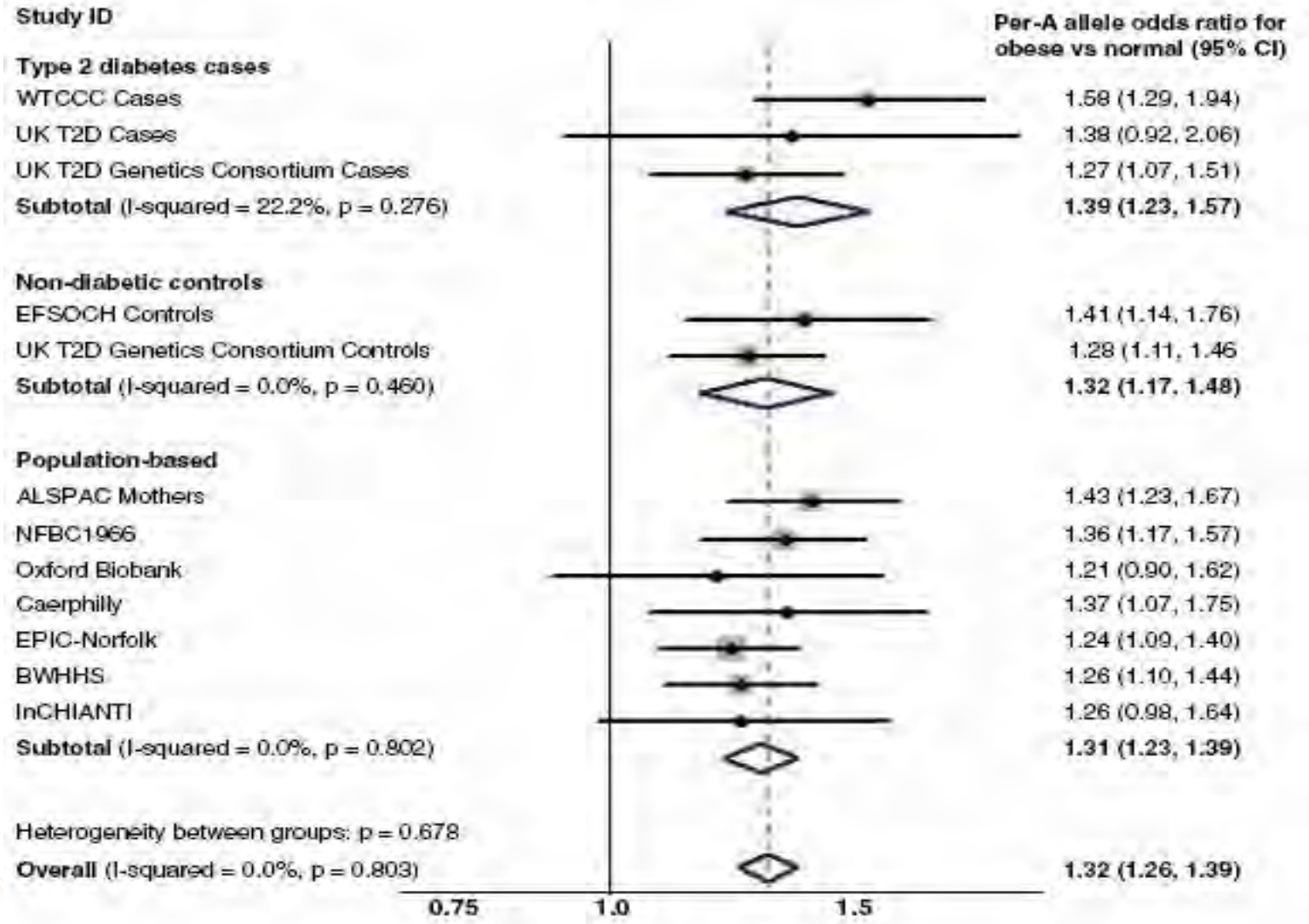
- Population stratification via EIGENSTRAT
  - SAS is very handy since a single put statement is sufficient to generate the output.
- Collaborative (e.g. height) and consortium work (GIANT)
  - On the UK side, this is mainly involved with IMPUTE/SNPTEST, with inputs on strand, standard error, quantitative traits, outputs.
  - This facilitates meta-analysis considerably.

The first obesity gene

A Common Variant in the *FTO* Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity

Timothy M. Frayling,[1,2]* Nicholas J. Timpson,[3,4]* Michael N. Weedon,[1,2]* Eleftheria Zeggini,[3,5]* Rachel M. Freathy,[1,2] Cecilia M. Lindgren,[3,5] John R. B. Perry,[1,2] Katherine S. Elliott,[3] Hana Lango,[1,2] Nigel W. Rayner,[3,5] Beverley Shields,[2] Lorna W. Harries,[2] Jeffrey C. Barrett,[3] Sian Ellard,[2,6] Christopher J. Groves,[5] Bridget Knight,[2] Ann-Marie Patch,[2,6] Andrew R. Ness,[7] Shah Ebrahim,[8] Debbie A. Lawlor,[9] Susan M. Ring,[9] Yoav Ben-Shlomo,[9] Marjo-Riitta Jarvelin,[10,11] Ulla Sovio,[10,11] Amanda J. Bennett,[5] David Melzer,[1,12] Luigi Ferrucci,[13] Ruth J. F. Loos,[14] Inês Barroso,[15] Nicholas J. Wareham,[14] Fredrik Karpe,[5] Katharine R. Owen,[5] Lon R. Cardon,[3] Mark Walker,[16] Graham A. Hitman,[17] Colin N. A. Palmer,[18] Alex S. F. Doney,[19] Andrew D. Morris,[19] George Davey Smith,[4] The Wellcome Trust Case Control Consortium,† Andrew T. Hattersley,[1,2]‡§ Mark I. McCarthy[3,5]‡

Meta-analysis for odds of obesity



LDL

LDL-cholesterol concentrations: a genome-wide association study

Manjinder S Sandhu, Dawn M Waterworth*, Sally L Debenham*, Eleanor Wheeler, Konstantinos Papadakis, Jing Hua Zhao, Kijoung Song, Xin Yuan, Toby Johnson, Sofie Ashford, Michael Inouye, Robert Luben, Matthew Sims, David Hadley, Wendy McArdle, Philip Barter, Y Antero Kesäniemi, Robert W Mahley, Ruth McPherson, Scott M Grundy, Wellcome Trust Case Control Consortium†, Sheila A Bingham, Kay-Tee Khaw, Ruth J F Loos, Gérard Waeber, Inês Barroso, David P Strachan, Panagiotis Deloukas, Peter Vollenweider, Nicholas J Wareham, Vincent Mooser*

Height

Genome-wide association analysis identifies 20 loci that influence adult height

Michael N Weedon[1,2,23], Hana Lango[1,2,23], Cecilia M Lindgren[3,4], Chris Wallace[5], David M Evans[6], Massimo Mangino[7], Rachel M Freathy[1,2], John R B Perry[1,2], Suzanne Stevens[7], Alistair S Hall[8], Nilesh J Samani[7], Beverly Shields[2], Inga Prokopenko[3,4], Martin Farrall[9], Anna Dominiczak[10], Diabetes Genetics Initiative[21], The Wellcome Trust Case Control Consortium[21], Toby Johnson[11–13], Sven Bergmann[11,12], Jacques S Beckmann[11,14], Peter Vollenweider[15], Dawn M Waterworth[16], Vincent Mooser[16], Colin N A Palmer[17], Andrew D Morris[18], Willem H Ouwehand[19,20], Cambridge GEM Consortium[22], Mark Caulfield[5], Patricia B Munroe[5], Andrew T Hattersley[1,2], Mark I McCarthy[3,4] & Timothy M Frayling[1,2]

Adult height is a model polygenic trait, but there has been limited success in identifying the genes underlying its normal variation. To identify genetic variants influencing adult human height, we used genome-wide association data from 13,665 individuals and genotyped 39 variants in an additional 16,482 samples. We identified 20 variants associated with adult height ($P < 5 \times 10^{-7}$, with 10 reaching $P < 1 \times 10^{-10}$). Combined, the 20 SNPs explain ~3% of height variation, with a ~5 cm difference between the 6.2% of people with 17 or fewer 'tall' alleles compared to the 5.5% with 27 or more 'tall' alleles. The loci we identified implicate genes in Hedgehog signaling (*IHH, HHIP, PTCH1*), extracellular matrix (*EFEMP1, ADAMTSL3, ACAN*) and cancer (*CDK6, HMGA2, DLEU7*) pathways, and provide new insights into human growth and developmental processes. Finally, our results provide insights into the genetic architecture of a classic quantitative trait.

BMI/obesity

Common variants near *MC4R* are associated with fat mass, weight and risk of obesity

Ruth J F Loos[*,1,2,73], Cecilia M Lindgren[3,4,73], Shengxu Li[1,2,73], Eleanor Wheeler[5], Jing Hua Zhao[1,2], Inga Prokopenko[3,4], Michael Inouye[5], Rachel M Freathy[6,7], Antony P Attwood[5,8], Jacques S Beckmann[9,10], Sonja I Berndt[11], The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial[71], Sven Bergmann[9,12], Amanda J Bennett[3,4], Sheila A Bingham[13], Murielle Bochud[14], Morris Brown[15], Stéphane Cauchi[16], John M Connell[17], Cyrus Cooper[18], George Davey Smith[19], Ian Day[18], Christian Dina[16], Subhajyoti De[20], Emmanouil T Dermitzakis[5], Alex S F Doney[21], Katherine S Elliott[3], Paul Elliott[22,23], David M Evans[3,19], I Sadaf Farooqi[2,24], Philippe Froguel[16,25], Jilur Ghori[5], Christopher J Groves[3,4], Rhian Gwilliam[5], David Hadley[26], Alistair S Hall[27], Andrew T Hattersley[6,7], Johannes Hebebrand[28], Iris M Heid[29,30], KORA[71], Blanca Herrera[3,4], Anke Hinney[28], Sarah E Hunt[5], Marjo-Riitta Jarvelin[22,23,31], Toby Johnson[9,12,14], Jennifer D M Jolley[8], Fredrik Karpe[4], Andrew Keniry[5], Kay-Tee Khaw[32], Robert N Luben[32], Massimo Mangino[33], Jonathan Marchini[34], Wendy L McArdle[35], Ralph McGinnis[5], David Meyre[16], Patricia B Munroe[36], Andrew D Morris[21], Andrew R Ness[37], Matthew J Neville[4], Alexandra C Nica[5], Ken K Ong[1,2], Stephen O'Rahilly[2,24], Katharine R Owen[4], Colin N A Palmer[38], Konstantinos Papadakis[26], Simon Potter[5], Anneli Pouta[31,39], Lu Qi[40], Nurses' Health Study[71], Joshua C Randall[3,4], Nigel W Rayner[3,4], Susan M Ring[35], Manjinder S Sandhu[1,32], André Scherag[41], Matthew A Sims[1,2], Kijoung Song[42], Nicole Soranzo[5], Elizabeth K Speliotes[43,44], Diabetes Genetics Initiative[71], Holly E Syddall[18], Sarah A Teichmann[20], Nicholas J Timpson[3,19], Jonathan H Tobias[45], Manuela Uda[46], The SardiNIA Study[71], Carla I Ganz Vogel[28], Chris Wallace[36], Dawn M Waterworth[42], Michael N Weedon[6,7], The Wellcome Trust Case Control Consortium[72], Cristen J Willer[47], FUSION[71], Vicki L Wraight[2,24], Xin Yuan[42], Eleftheria Zeggini[3], Joel N Hirschhorn[44,48–51], David P Strachan[26], Willem H Ouwehand[8], Mark J Caulfield[36], Nilesh J Samani[33], Timothy M Frayling[6,7], Peter Vollenweider[52], Gerard Waeber[52], Vincent Mooser[42], Panos Deloukas[5], Mark I McCarthy[3,4,73], Nicholas J Wareham[1,2,73] & Inês Barroso[5,73]

Further on BMI

Six new loci associated with body mass index highlight a neuronal influence on body weight regulation

*Cristen J Willer[1,77,78], Elizabeth K Speliotes[2,3,77,78], Ruth J F Loos[4,5,77,78], Shengxu Li[4,5,77,78], Cecilia M Lindgren[6,78], Iris M Heid[7,78], Sonja I Berndt[8], Amanda L Elliott[9,10], Anne U Jackson[1], Claudia Lamina[7], Guillaume Lettre[9,11], Noha Lim[12], Helen N Lyon[3,11], Steven A McCarroll[9,10], Konstantinos Papadakis[13], Lu Qi[14,15], Joshua C Randall[6], Rosa Maria Roccasecca[16], Serena Sanna[17], Paul Scheet[18], Michael N Weedon[19], Eleanor Wheeler[16], Jing Hua Zhao[4,5], Leonie C Jacobs[20], Inga Prokopenko[6,21], Nicole Soranzo[16,22], Toshiko Tanaka[23], Nicholas J Timpson[24], Peter Almgren[25], Amanda Bennett[26], Richard N Bergman[27], Sheila A Bingham[28,29], Lori L Bonnycastle[30], Morris Brown[31], Noël P Burtt[9], Peter Chines[30], Lachlan Coin[32], Francis S Collins[30], John M Connell[33], Cyrus Cooper[34], George Davey Smith[24], Elaine M Dennison[34], Parimal Deodhar[30], Paul Elliott[32], Michael R Erdos[30], Karol Estrada[20], David M Evans[24], Lauren Gianniny[9], Christian Gieger[7], Christopher J Gillson[4,5], Candace Guiducci[9], Rachel Hackett[9], David Hadley[13], Alistair S Hall[35], Aki S Havulinna[36], Johannes Hebebrand[37], Albert Hofman[38], Bo Isomaa[39], Kevin B Jacobs[40], Toby Johnson[41–43], Pekka Jousilahti[36], Zorica Jovanovic[5,44], Kay-Tee Khaw[45], Peter Kraft[46], Mikko Kuokkanen[9,47], Johanna Kuusisto[48], Jaana Laitinen[49], Edward G Lakatta[50], Jian'an Luan[4,5], Robert N Luben[45], Massimo Mangino[51], Wendy L McArdle[52], Thomas Meitinger[53,54], Antonella Mulas[17], Patricia B Munroe[55], Narisu Narisu[30], Andrew R Ness[56], Kate Northstone[52], Stephen O'Rahilly[5,44], Carolin Purmann[5,44], Matthew G Rees[30], Martin Ridderstråle[57], Susan M Ring[52], Fernando Rivadeneira[20,38], Aimo Ruokonen[58], Manjinder S Sandhu[4,45], Jouko Saramies[59], Laura J Scott[1], Angelo Scuteri[60], Kaisa Silander[47], Matthew A Sims[4,5], Kijoung Song[12], Jonathan Stephens[61], Suzanne Stevens[51], Heather M Stringham[1], Y C Loraine Tung[5,44], Timo T Valle[62], Cornelia M Van Duijn[38], Karani S Vimaleswaran[4,5], Peter Vollenweider[63], Gerard Waeber[63], Chris Wallace[55], Richard M Watanabe[64], Dawn M Waterworth[12], Nicholas Watkins[61], The Wellcome Trust Case Control Consortium[76], Jacqueline C M Witteman[38], Eleftheria Zeggini[6], Guangju Zhai[22], M Carola Zillikens[20], David Altshuler[9,10], Mark J Caulfield[55], Stephen J Chanock[8], I Sadaf Farooqi[5,44], Luigi Ferrucci[23], Jack M Guralnik[65], Andrew T Hattersley[66], Frank B Hu[14,15], Marjo-Riitta Jarvelin[32], Markku Laakso[48], Vincent Mooser[12], Ken K Ong[4,5], Willem H Ouwehand[16,61], Veikko Salomaa[36], Nilesh J Samani[51], Timothy D Spector[22], Tiinamaija Tuomi[67,68], Jaakko Tuomilehto[62], Manuela Uda[17], André G Uitterlinden[20,38], Nicholas J Wareham[4,5], Panagiotis Deloukas[16], Timothy M Frayling[19], Leif C Groop[25,69], Richard B Hayes[8], David J Hunter[9,14,15,46], Karen L Mohlke[70], Leena Peltonen[9,16,71], David Schlessinger[72], David P Strachan[13], H-Erich Wichmann[7,73], Mark I McCarthy[6,21,74,78,79], Michael Boehnke[1,78,79], Inês Barroso[16,78,79], Gonçalo R Abecasis[18,78,79] & Joel N Hirschhorn[3,11,75,78,79] for the GIANT Consortium[80]

Reflection on the study design

	Study 1 (EPIC-Norfolk subcohort) n=2269		Study 2 (EPIC-Norfolk obese set) n=1009		Study 3 (1958 British birth cohort) n=1375		Study 4 (CoLaus) n=5367		Study 5 (GEMS study) n=1665	
	β coeff (SE)	p value	β coeff (SE)	p value	β coeff (SE)	p value	β coeff (SE)	p value	β coeff (SE)	p value
rs4420638	0·24 (0·04)	$1·9\times10^{-9}$	0·14 (0·06)	0·02	0·25 (0·04)	$2·8\times10^{-9}$	0·05 (0·01)	$6·2\times10^{-12}$	0·04 (0·01)	$5·6\times10^{-3}$
rs599839	−0·15 (0·04)	$5·8\times10^{-5}$	−0·23 (0·06)	$7·6\times10^{-5}$	−0·14 (0·04)	$4·3\times10^{-4}$	−0·04 (0·01)	$1·6\times10^{-7}$	−0·06 (0·01)	$2·0\times10^{-5}$
rs4970834	−0·13 (0·04)	$1·1\times10^{-3}$	−0·18 (0·06)	$5·5\times10^{-3}$	−0·11 (0·04)	0·01	−0·04 (0·01)	$1·9\times10^{-6}$	−0·04 (0·01)	$2·8\times10^{-3}$
rs562338	−0·17 (0·04)	$6·0\times10^{-6}$	−0·11 (0·06)	0·07	−0·18 (0·05)	$1·1\times10^{-4}$	−0·03 (0·01)	$2·7\times10^{-6}$	−0·02 (0·01)	0·18
rs7575840	0·15 (0·03)	$6·3\times10^{-6}$	0·15 (0·05)	$2·4\times10^{-3}$	0·04 (0·04)	0·26	0·03 (0·01)	$1·9\times10^{-6}$	0·02 (0·01)	0·13
rs478442	−0·16 (0·04)	$2·1\times10^{-5}$	−0·07 (0·06)	0·25	−0·16 (0·04)	$3·6\times10^{-4}$	−0·03 (0·01)	$2·7\times10^{-5}$	−0·02 (0·01)	0·06
rs4591370	−0·17 (0·04)	$7·7\times10^{-6}$	−0·06 (0·06)	0·28	−0·16 (0·04)	$4·2\times10^{-4}$	−0·03 (0·01)	$3·2\times10^{-5}$	−0·02 (0·01)	0·06
rs4560142	−0·16 (0·04)	$1·6\times10^{-5}$	−0·06 (0·06)	0·27	−0·16 (0·04)	$4·2\times10^{-4}$	−0·03 (0·01)	$3·5\times10^{-5}$	−0·03 (0·01)	0·05
rs576203	−0·16 (0·04)	$1·2\times10^{-5}$	−0·07 (0·06)	0·25	−0·16 (0·04)	$3·5\times10^{-4}$	−0·03 (0·01)	$3·5\times10^{-5}$	−0·02 (0·01)	0·06
rs506585	−0·16 (0·04)	$1·7\times10^{-5}$	−0·06 (0·06)	0·31	−0·16 (0·04)	$3·5\times10^{-4}$	−0·03 (0·01)	$4·2\times10^{-5}$	−0·03 (0·01)	0·05
rs488507	−0·14 (0·04)	$1·3\times10^{-4}$	−0·07 (0·06)	0·25	−0·16 (0·04)	$3·3\times10^{-4}$	−0·03 (0·01)	$3·4\times10^{-5}$	−0·02 (0·01)	0·07
rs538928	−0·16 (0·04)	$5·0\times10^{-5}$	−0·01 (0·06)	0·92	−0·16 (0·04)	$3·5\times10^{-4}$	−0·03 (0·01)	$3·6\times10^{-5}$	−0·02 (0·01)	0·05
rs10402271	0·04 (0·03)	0·17	0·11 (0·05)	0·02	0·12 (0·04)	$7·5\times10^{-4}$	0·02 (0·01)	$5·2\times10^{-4}$	0·04 (0·01)	$8·3\times10^{-4}$
rs693	−0·12 (0·03)	$1·3\times10^{-4}$	−0·07 (0·05)	0·15	−0·06 (0·03)	0·06	−0·03 (0·01)	$1·0\times10^{-5}$	−0·02 (0·01)	0·16

Webtable 3: Associations between Affymetrix SNPs with a combined p value of $<1·0\times10^{-7}$ and circulating concentrations of LDL cholesterol in independent study populations

Our best practice

- Linux clusters are now ready for comphensive analyses.
- Linux/awk script is light and appears to be more transparent than Perl, Java which is more professional.
- awk proves very useful and can be transformed to Perl. In fact, any statistical package which processes data elements would be less efficient. An example is the transformation of long, wide, transposed format noted earlier.
- They call C/C++ programs such as IMPUTE/SNPTEST.
- SAS is still useful for data preparation, and in a sense less professional than DBMS such as Oracle but enjoys a large user community and has facility for data analysis.
- SAS 9.2 PROTO procedure is yet to be explored.

The transpose data format

rs17782313	TT	CT	TT	TT	TT	TT	TT	CC
rs8097644	CC	CC	AC	CC	CC	CC	CC	CC
rs9947403	CC	TT	CC	CC	CC	CC	CC	TT
rs639407	AA	GG	AA	AA	AA	AA	AA	GG
rs11665563	CC	CT	CC	CC	CC	CC	CC	TT
rs11663816	TT	CT	TT	TT	TT	TT	TT	CC
rs619662	GG	AA	AG	GG	GG	GG	GG	AA
rs727406	GG	GG	GT	GG	0	0	GG	GG
rs8089366	GG	GT	GG	GG	GG	GG	GG	TT
rs11152217	GG	GT	GG	GG	GG	GG	GG	GG
rs9955666	GG	AG	GG	GG	GG	GG	GG	AA
rs17700633	GG	AG	GG	GG	GG	GG	AG	AA
rs9946888	TT	CT	CT	CT	TT	CT	CT	TT
rs9961245	CC	CT	CT	CT	CC	CT	CT	CC
rs17066774	GG	GG	GG	GG	GG	GG	GG	GG

Data generation in SAS

```
data long (keep=&snpid id &vlist a1a2 add n);
      set data; fid=open("data");
      length id $11. add 3. a1a2 $3.; format add 1.;
      set map point=_n_;
      n=0;
      do col=2 to attrn(fid,"nvars");
          iid=col-1;
          set &trait (keep=&vlist) point=iid;
          if &inc=1 then do;
            id=varname(fid,col); a1a2=vvaluex(id);  add=.;
            if a1a2 ne "   " then do;
              a1=substr(a1a2,1,1); a2=substr(a1a2,3,1);
              add=(a1=b)+(a2=b);
              n+1;
            end; output;
          end;
      end; rc=close(fid);
run;
```

Meta-analysis (fixed-effects)

```
data test;
        input studyid lor est;
        col=_n_; row=_n_;
        value=est;
Cards;
... data for 15 studies ...
run;
proc mixed method = ml data=test;
         class studyid;
         model lor = / s cl;
         repeated / group = studyid;
         parms / parmsdata=test eqcons=1 to 15;
run;
```

Meta-analysis (random-effects)

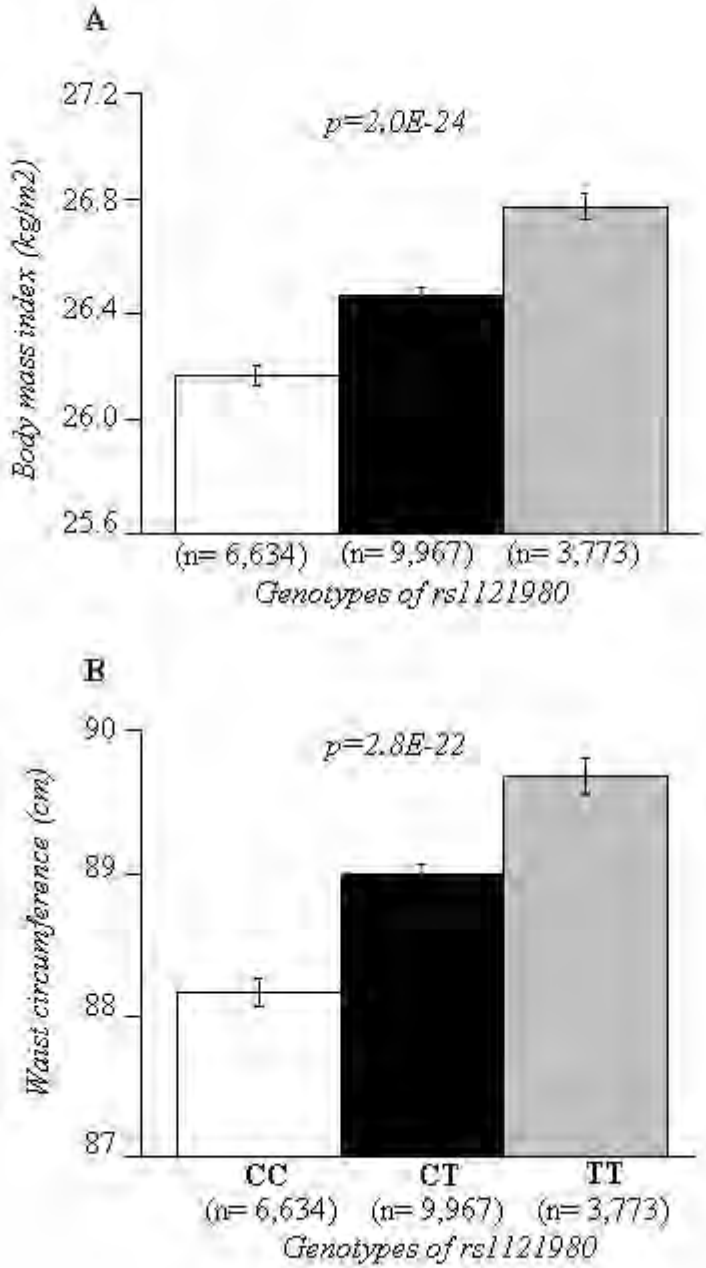
```
proc mixed data=test covtest; /*no specification of 15*/
        class studyid;
        model lor = / s cl outp=predp outpm=predm;
        repeated diag / r;
        random studyid / g gdata = test s v;
        ods output CovParms=cp G=G R=R V=V
                        SolutionF=SF SolutionR=SR;
run;
data predp;
      set predp; pvalue=probnorm(resid/stderrpred);
run;
data predm;
      set predm;pvalue=probnorm(resid/stderrpred);
run;
```

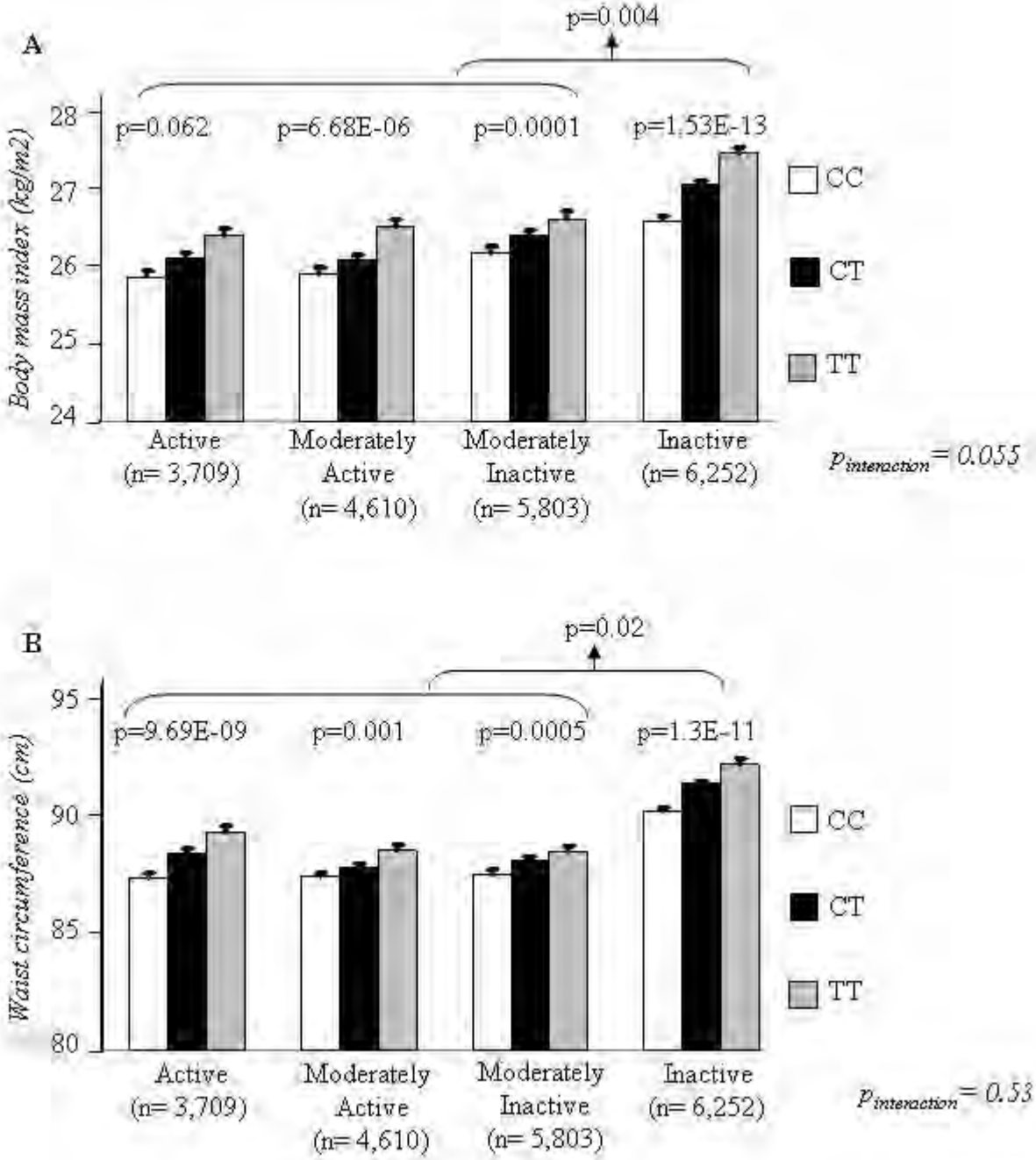
SAS/Genetics and SAS/STAT

```
proc allele data=long genocol;
      by rsn notsorted;
      var a1a2;
      ods output markersumm=ms allelefreq=out.af;
run;
proc reg data=long;
      by rsn notsorted;
      ods output parameterestimates=bmipm;
      model bmi = age add / b stb;
quit;
proc logistic data=long descending;
      by rsn notsorted;
      ods output parameterestimates=obpm CLOddsPL=obclpm;
      model obesity = age add / expb clodds=pl;
run;
```

FTO/physical activity--BMI/WC association

- *FTO* variant, rs1121980, was genotyped in 20,374 participants (39-79 years) from the EPIC-Norfolk Study. Physical activity (PA) was assessed by a validated self-reported questionnaire. The interaction between rs1121980 and PA on BMI and waist circumference (WC) was examined by including the interaction term in mixed effect models.
- Our results show that PA attenuates the effect of *FTO* rs1121980 genotype on BMI and WC.


A
27.2
26.8
26.4
26.0
25.6
p=2.0E-24
Body mass index (kg/m2)
(n= 6,634)
(n= 9,967)
(n= 3,773)
Genotypes of rs1121980
B
90
89
88
87
p=2.8E-22
Waist circumference (cm)
CC
CT
TT
(n= 6,634)
(n= 9,967)
(n= 3,773)
Genotypes of rs1121980


A
p=0.004
p=0.062
p=6.68E-06
p=0.0001
p=1.53E-13
Body mass index (kg/m2)
28
27
26
25
24
CC
CT
TT
Active (n= 3,709)
Moderately Active (n= 4,610)
Moderately Inactive (n= 5,803)
Inactive (n= 6,252)
$p_{interaction}$ = 0.055
B
p=0.02
p=9.69E-09
p=0.001
p=0.0005
p=1.3E-11
Waist circumference (cm)
95
90
85
80
CC
CT
TT
Active (n= 3,709)
Moderately Active (n= 4,610)
Moderately Inactive (n= 5,803)
Inactive (n= 6,252)
$p_{interaction}$ = 0.53

References

- Bodmer W, Bonilla C. Nat Genet 2008;40:695-701
- EPIC: http://epic.iarc.fr/
- EPIC-Norfolk: http://www.srl.cam.ac.uk/epic
- Long AD et al.. Science 1997; 275:1328
- Loos R et al. Nat Genet 2008; 40:468-75
- Prentice RL. Biometrika 1986;73:1-11
- Risch N, Merkangas K (1996) Science 1997;273:1516-7
- Sandhu MS et al. Lancet 2008; 371:483-91
- Santhanakrishnan V. Am J Clin Nutr (in press)
- Weedon MN et al. 2008;40:575-83
- Willer et al. Nat Genet 2009; 41:25-34
- Zhao JH. J Stat Soft 2007;23(8):1-18
- Zhao JH et al. CCIS 2007;2:781-90

Institute of Metabolic Science
MRC | Epidemiology Unit

Genome-wide Association Studies

Association testing

Jing Hua Zhao

Topics

- Elements of analysis
- R packages
- Example: *MC4R* and obesity

Elements of analysis

- Quality control: call rates, Hardy-Weinberg equilibrium and minor allele frequencies and others such as clustering of genotypes, relatedness and population stratification.
- Test of associations
  - often through linear regression for continuous trait, and through logistic regression for binary, the proportion of variance explained for LR is measured through R^2 while the score statistic under additive model is equivalent to the Armitage trend test.
  - Genotype imputation: mostly often through HapMap CEU sample, involving ~2.5 million SNPs
  - Graphical presentation
- Interpretation, replication
- Report of findings

Graphics and association plots

- Plot of summary statistics
- Pedigree-drawing
- LD plot
- Q-Q plot -- contrasting observed versus expected log-p values
- Manhattan plot -- distribution of genome-wide p values
- Regional association plot -- including recombination, contribution from imputed SNPs and top hits from consortium meta-analysis

Basic R packages

- genetics
- haplo.stats
- gap
- kinship, multic, pedigree, identity

- See CRAN task view for Genetics (http://cran.r-project.org/web/views/Genetics.html), as with an earlier review on the motivation for analysis with R statistical and computational environment (Zhao & Tan. *Hum Genomics* 2006;2:258-65)
- It also refers to Rgenetics projects whose packages are now available from http://www.bioconductor.org.

R packages for GWAS

- SNPassoc
- GenABEL
- P2BAT
- snpMatrix

- While the first three are available from CRAN, snpMatrix is available from BioConductor.
- Other packages include, multtest, meta, rmeta, CAMAN, qvalue, ROCR.

Installation and use

- BioConductor

  http://bioconductor.org/packages/2.3/bioc/html/snpMatrix.html

  > source("http://bioconductor.org/biocLite.R")

  > biocLite("snpMatrix")
- CRAN

  http://cran.r-project.org/web/packages/index.html

> setRepositories()

> install.packages(“snpMatrix”)

> library(snpMatrix)

Example – *MC4R* SNPs and BMI

- To make a smooth exposition we use our study of SNPs near *MC4R* and body mass index as reported by Loos et al. *Nat Genet* 2008;40:768-775.
- The *MC4R* gene is located on chromosome 18 and we will focus on SNPs rs17782313 and rs17700633 at positions 56002077 and 57000671 according to NCBI build 35, all genotypes being on forward strand.
- These were based on 3850 population-based individuals at stage 1 of the case-cohort study from which 3552 individuals remained after quality controls.
- We will run through SNPassoc, snpMatrix and GenABEL packages on data as contained in files mc4r.ped, mc4r.map and mc4r.csv

SNPassoc: data input

```
library(SNPassoc)
map <- read.table("mc4r.map",sep="\t",as.is=TRUE)
info <- data.frame(snp=map[2],chr=map[2],pos=map[4])
ped <- read.table("mc4r.ped",sep="\t",as.is=TRUE)
names(ped) <- c(paste("v",1:6,sep=""),map[,2])
pheno <-
   read.csv("mc4r.csv",sep="\t",skip=11,header=TRUE,as.is=TR
   UE)
is.cohort <- pheno$cohort==1
cohort <- subset(pheno,is.cohort)
snp <- ped[,-c(1:6)][is.cohort,]
snps <- dim(snp)[2]
for(i in 1:snps)
{
   substr(snp[,i],2,2) <- "/"; empty <- (snp[,i]=="0/0");
   snp[empty,i] <- NA
}
```

SNPassoc: analysis

```
mc4r <- setupSNP(snp,1:snps,sep="/",sort=TRUE,info=info)
summary(mc4r)
plot(mc4r$rs17782313)
plot(mc4r$rs17700633,type=pie)
hwe <- tableHWE(mc4r)
mc4r.ld <- LD(mc4r)
summary(mc4r.ld)
mc4r.ld$"R^2"
attach(cohort)
association(bmi ~ sex+age+rs17782313,data=mc4r)
wga <- WGassociation(bmi ~ sex+age+1,model="log-
   add",data=mc4r)
png("mc4r.png")
qqpval(wga$"log-additive")
dev.off()
```

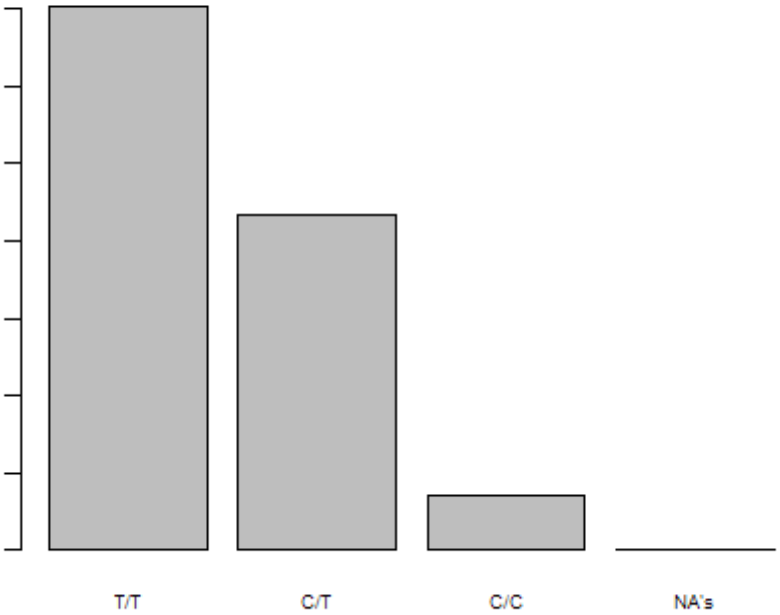
Summary statistics for two SNPs

mc4r$rs17782313

	frequency	percentage
C	1150	23.81
T	3680	76.19
NA's	4	NA

	frequency	percentage
T/T	1405	58.18
C/T	870	36.02
C/C	140	5.80
NA's	2	NA

HWE (pvalue): 0.736476

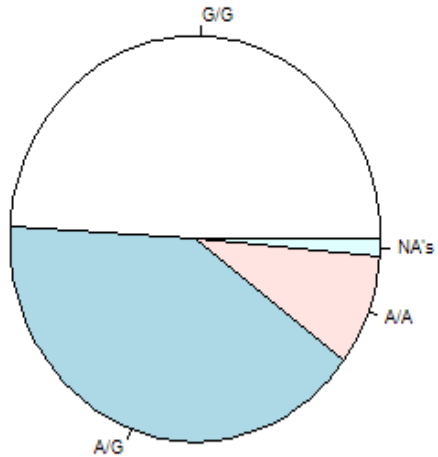


mc4r$rs17700633

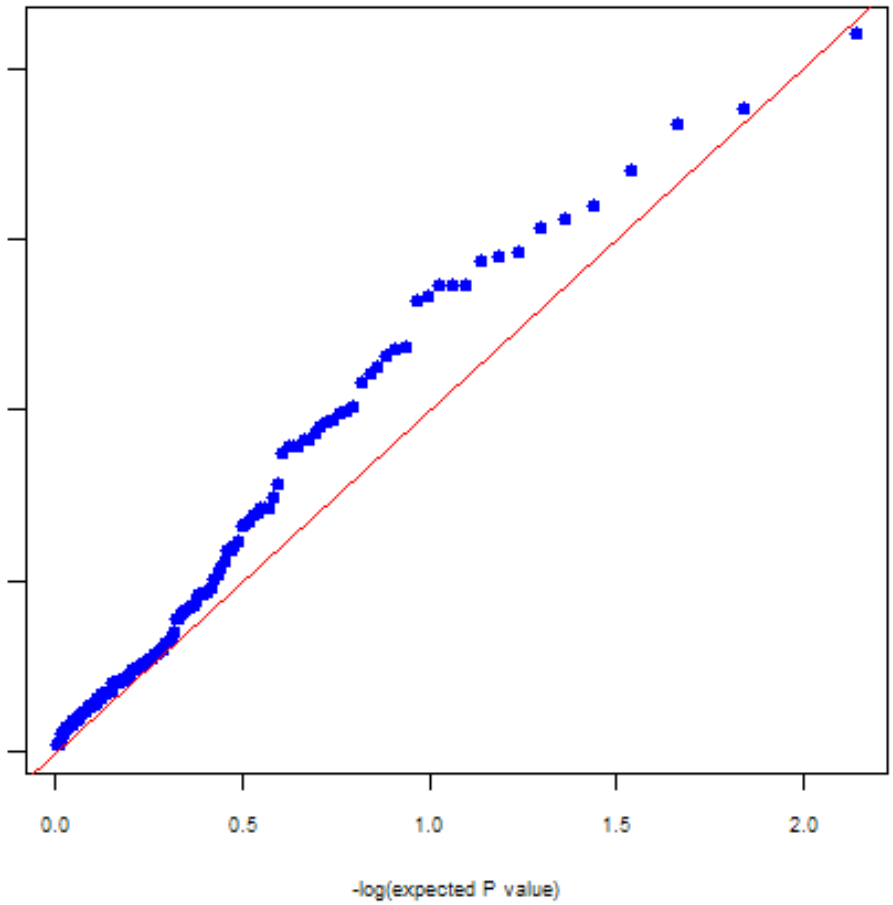
	frequency	percentage
A	1411	29.62
G	3353	70.38
NA's	70	NA

	frequency	percentage
G/G	1183	49.66
A/G	987	41.44
A/A	212	8.90
NA's	35	NA

HWE (pvalue): 0.768142



SNPassoc: Q-Q plot



Adjustment for multiple testing

```
Bonferroni.sig(wga, model="log-add",alpha=0.05)
library(qvalue)
q <- qvalue(p)
plot(q)
library(multtest)
adj <-
   c("Bonferroni","Holm","Hochberg","SidakSS","SidakSD",
   "BH","BY")
mt <- mt.rawp2adjp(p,adj)
mt.reject(cbind(mt$rawp,mt$adjp),seq(0,0.1,0.001))$r
```

SNPassoc: comments

- SNPassoc is essentially designed for dealing with unrelated individuals but with considerable enhancements from genetics and haplo.stats.
- It implements permutation tests for binary traits through scanWGassociation(,nperm=) and permTest()
- It is possible to conduct gene-gene interaction:
  - mc4r.ip <- interactionPval(bmi~sex+age,data=mc4r,model="log-add")
  - plot(mc4r.ip)
- We got a very good feel of the kind of analysis it may involve and this is a very simple example.

haplo.stats: haplotype analysis

```
library(haplo.stats)
mc4r.map <- read.table("mc4r.map",as.is=TRUE)
snps <- mc4r.map[,2]
M <- length(snps)
a1 <- sprintf("%s%s",snps,rep(".a1",M))
a2 <- sprintf("%s%s",snps,rep(".a2",M))
a1a2 <- c(a1,a2)
for(i in 1:M) {a1a2[2*i-1] <- a1[i];a1a2[2*i] <- a2[i]}
mc4r <-
    read.table("mc4r.ped",col.names=c(paste("v",1:6,sep=""),a1a2))

pheno <- read.csv("mc4r.csv",sep="\t",skip=11)
cohort <- subset(pheno,cohort==1)
attach(cohort)
mc4r.12 <-
    haplo.score(bmi,mc4r[id,7:30],x.adj=sex+age,locus.label=snps[1:12
    ])
```

haplo.stats: haplo.glm

```
mc4r.geno <-
   setupGeno(mc4r[id,7:dim(mc4r)[2]],locus.label=snps)
attr(mc4r.geno,"unique.alleles")[1:12]
mc4r.12 <-
   haplo.score(bmi,mc4r.geno[,1:24],x.adj=sex+age,locus.label
   =snps[1:12])


mc4r.data <- data.frame(geno=mc4r.geno,cohort)
mc4r.gauss <- haplo.glm(bmi ~ sex + age + geno, family =
   gaussian, na.action="na.geno.keep",
                  allele.lev=attributes(geno)$unique.alleles,
                  data=mc4r.data, locus.label=snps,
                  control =
   haplo.glm.control(haplo.freq.min=0.02))
mc4r.gauss
detach(cohort)
```

- haplo.glm is considerably slower but it is among the few facilities for GEI analysis

snpMatrix: SNP data

```
library(snpMatrix)
mc4r <- read.snps.pedfile("mc4r.ped")
summary(mc4r)
head(mc4r$snp.support)
head(mc4r$subject.support)
# quality controls
mc4r.qc <- summary(mc4r$snp.data)
head(mc4r.qc)
mc4r.ld <- ld.snp(mc4r$snp.data)
plot.snp.dprime(mc4r.ld,"mc4r.eps",scheme="rsq")
# ps2pdf mc4r.eps
# xpdf mc4r.pdf
# LD(rs17782313, rs17700633)
mc4r$snp.support[c(1,12),]
pair.result.ld.snp(mc4r$snp.data,1,12)
```

PCA and identity-by-state analysis

```
# PCA
mc4r.xxt <- xxt(mc4r$snp.data,correct.for.missing=TRUE)
mc4r.pc <- eigen(mc4r.xxt, symmetric=TRUE)
loadings <- snp.cor(mc4r$snp.data, mc4r.pc$vectors[,1:10])
# identity-by-state analysis
mc4r.ibs <- ibs.stats(mc4r$snp.data)
mc4r.count <- ibsCount(mc4r$snp.data)
mc4r.dist <- ibsDist(mc4r.dist)
mc4r.clust <0 hclus(mc4r.dist)
plot(mc4r.clust)
```

- Note this is based on XX^T, in an order of N^2, where X and N are the genotype data matrix and number of individuals (see vignette for details).

Phenotype data and case-control analysis

```
# Phenotype data
pheno <- read.csv("mc4r.csv",skip=11,sep="\t")
pheno$cc<-ifelse(pheno$bmi>=30,1,0)
attach(pheno)

# Case-control analysis of all individuals
cc.test <- single.snp.tests(cc,snp.data=mc4r$snp.data)
summary(cc.test)
class(cc.test)
showClass("snp.tests.single")
chi.squared(cc.test,1) #qq.chisq(cc.test@chisq[,1])
```

Meta-analysis

```
# a meta-analysis
cc.test <-
   single.snp.tests(cc,snp.data=mc4r$snp.data,score=TRU
   E)
cc.test2 <- pool(cc.test,cc.test)
summary(cc.test2)
cc.test.sign <- effect.sign(cc.test)
table(cc.test.sign)
cc.test.sign[1:12]
cc.test.switch <- switch.alleles(cc.test,c(1,12))
effect.sign(cc.test.switch)[1:12]
```

OLS estimation and retrospective analysis

```
# ordinary least squares estimates
reg.rhs <-
    snp.rhs.tests(bmi~sex+age,family="gaussian",subset=(cohort==1),s
    np.data=mc4r$snp.data)
class(reg.rhs)
showClass("snp.tests.glm")
reg.rhs@df
qq.chisq(reg.rhs@chisq)
print(reg.rhs)

# retrospective models
reg.lhs <-
    snp.lhs.tests(mc4r$snp.data,~bmi,~sex+age,subset=(cohort==1))
class(reg.lhs)
showClass("snp.tests.glm")
reg.rhs@df
qq.chisq(reg.lhs@chisq,df=2)
```

Genotype imputation

- It is customarily to impute genotypes in a large study based on a small sample of fully-genotyped individuals, e.g., hapmap, so as to conduct association tests for large number of SNPs.
- It is also useful for meta-analysis of SNPs from different platforms such as Affymetrix 500K and Illumina 550K.
- As it is snpMatrix implements genotype imputation between sets of markers based on same individuals; more generally this involves genotypes from HapMap.

Hapmap and imputation

```
# ideally we would use 60/90 founders and a combination of hapmap
    CEU and our study sample
url.p1 <- "http://ftp.hapmap.org"
url.p2 <- "/genotypes/latest_ncbi_build35/fwd_strand/non-redundant/"
url.p3 <- "genotypes_chr18_CEU_r21a_nr_fwd.txt.gz"
hapmap <- paste(url.p1,url.p2,url.p3,sep="")
chr18 <- read.HapMap.data(hapmap)
summary(chr18)
sel <-
    row.names(chr18$snp.support)%in%row.names(mc4r$snp.support)
impute.from <- chr18$snp.data[,!sel]
impute.to <- mc4r$snp.data
pos.from <- chr18$snp.support$Position[!sel]
pos.to <- mc4r$snp.support$position
mc4r.imp <- snp.imputation(impute.from, impute.to, pos.from, pos.to)
summary(mc4r.imp)
plot(mc4r.imp)
```

snp.imputation: comments

- It is notable with the definition of snp.imputation that Given two set of SNPs typed in the same subjects, this function calculates regression equations which can be used to impute one set from the other in a subsequent sample.
- One we customarily use external data (e.g., available from HapMap, 1000 genomes or elsewhere) and our sample jointly, treating non-typed SNPs as missing.
- CRAN packages such as mice should facilitate this.

snpMatrix: comments

- snpMatrix has explicit treatment of data for chromosome X but that should be similar to method of autosome data in principle.
- It also provides some facilities for dealing with family data.
- The retrospective method would be more appropriate with data involving the kind of sample selection here.
- It is possible to take advantage of the S4 class facility as implemented in the package when coded genotypes are available from or to other sources, e.g.,
  - m1 <- new('snp.matrix',dm1)
  - m2 <- new('snp.matrix',dm2)
  - m <- snp.rbind(m1,m2)
  - write.snp.matrix(m,"m.dat")
- Please check for snpMatrix vignette for use of hexbin package.

GenABEL: data input

```
library(GenABEL)
convert.snp.ped("mc4r.ped","mc4r.map2","mc4r.out",strand="+
    ")
csv <- read.csv("mc4r.csv",skip=11,sep="\t",as.is=TRUE)
attach(csv)
csv2 <- data.frame(id,sex=2-sex,cohort,age,bmi,zbmi,rbmi)
write.table(csv2,"mc4r.csv2",sep=" ",row.names=FALSE)
mc4r <- load.gwaa.data(phe = "mc4r.csv2", gen = "mc4r.out",
    force = TRUE)
```

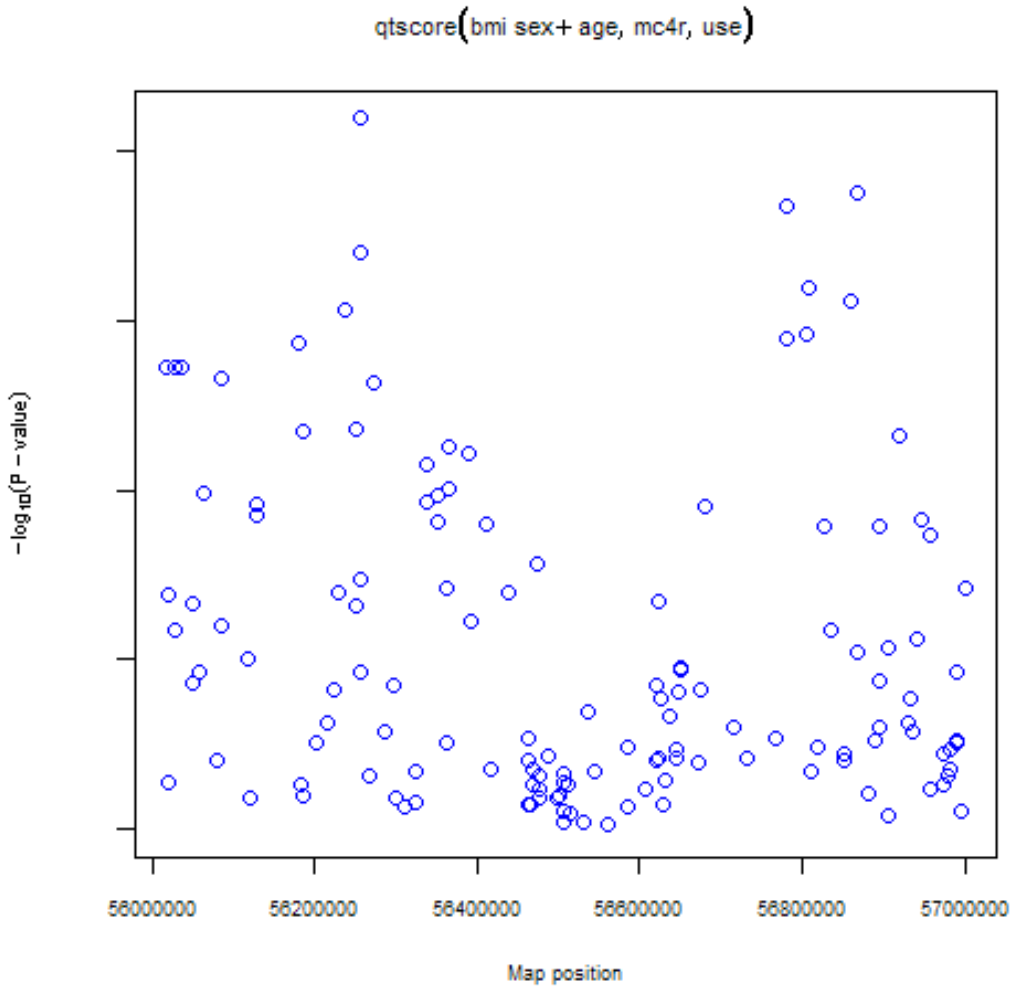
- Note that the map2 file now has three columns: chromosome, SNP names and positions. It also explicitly allow for strand. The addition of phenotypic information is via the load.gwaa.data, which requires specification of id and sex (0=female, 1=male) in a strictly way.

GenABEL: analysis

```
HWE.show(mc4r)
r2 <- r2fast(mc4r)
dp <- dprfast(mc4r)
rho <- rhofast(mc4r)
descriptives.trait(mc4r)
descriptives.marker(mc4r)
use <- csv$cohort==1
qt.bmi <- qtscore(bmi~sex+age,data=mc4r,idsubset=use)
plot(qt.bmi)
```

- However, as is shown here once the gwaa.data is defined a range of analyses can be rather straightforward.
- Again we only focus on the cohort sample (cohort==1).

GenABEL: scatter plot of p values



GRAMMAR

- It stands for genome-wide rapid association using mixed model and regression. The method first obtains residuals adjusted for family effects and subsequently analyzes the association between these residuals and genetic polymorphisms using least-squares methods. It can also involves selected polymorphism to be followed up with the full measured genotype analysis (Aulchenko et al. Genetics 2007).
- Initial model: $y_i = \mu + \Sigma_j \beta_j X_{ij} + G_i + e_i$

  We have the residuals $\hat{e}_i = y_i - (\hat{\mu} + \Sigma_j \hat{\beta}_j X_{ij} + \hat{G}_i) \equiv y_i *$
- Linear regression: $\hat{e}_i = \varphi + \gamma_i g_i + \varepsilon_i$
- Measured genotype model: $y_i = \mu + \gamma_i g_i + \Sigma_j \beta_j X_{ij} + G_i + e_i$

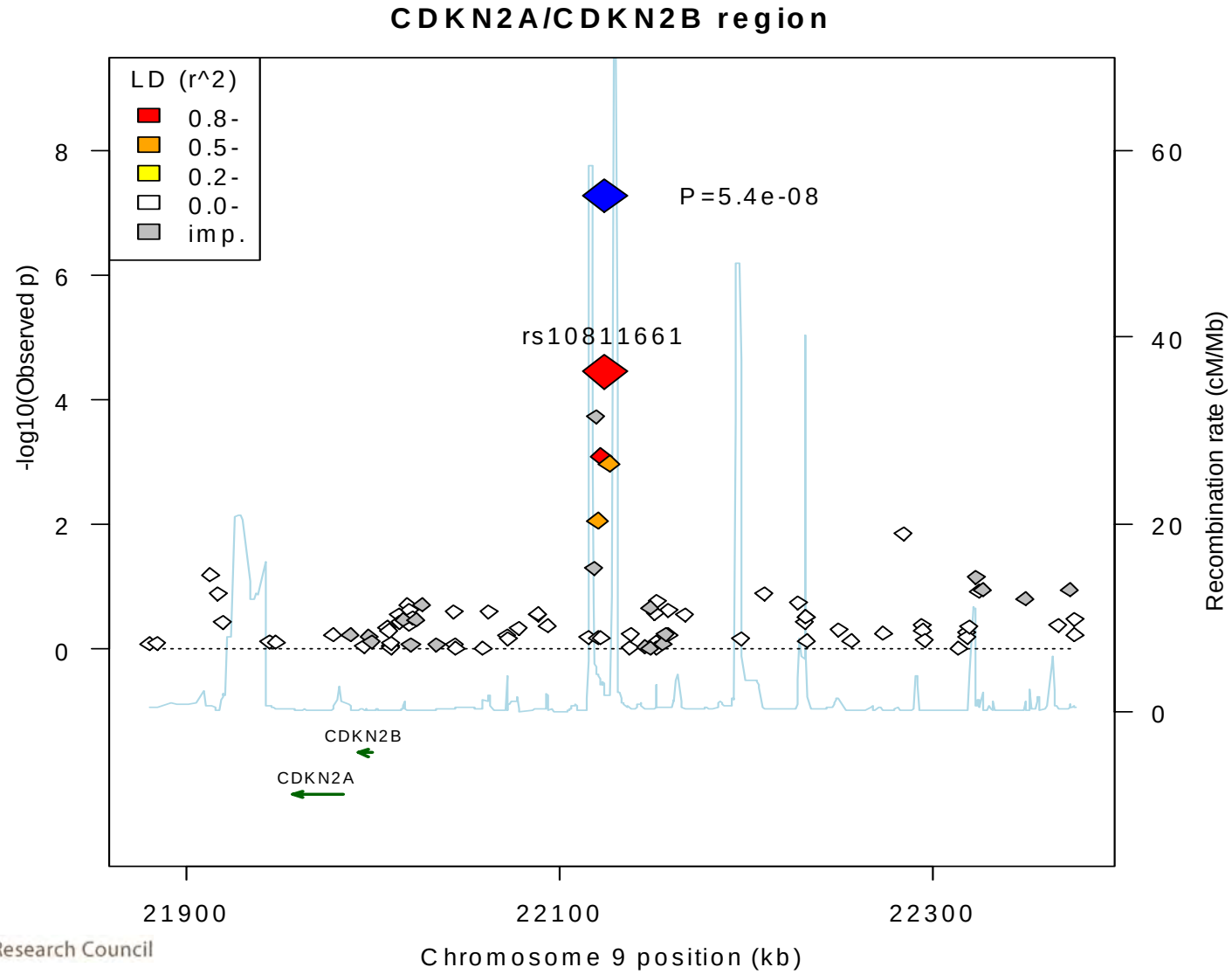
- The method adjusts for familial relationship, computationally fast, and ready to incorporate methods developed for “unrelated“ individuals in the second stage.

Manhattan and regional association plots

```
library(gap)
png("figures.png")
par(mfrow=c(2,1),mai=c(1,1,0.2,0.8),ps=7)
qqunif(test$np,bg="blue",bty="n",xlim=c(0,6),cex=0.02)
par(las=2)
mhtplot(test,usepos=TRUE,pch=21,colors=rep(c("blue","green"),
    11),cutoffs=c(4,5,6),cex=0.02)
dev.off()

asplot("rs10811661", "CDKN2A/CDKN2B region", "9",
    CDKNlocus, CDKNmap, CDKNgenes, 5.4e-8, c(3,6))
```

Association plot



Institute of Metabolic Science

MRC | Epidemiology Unit

Genome-wide Association Studies

Meta-analysis

Jing Hua Zhao

Topics

- Rationale
- Models
- Implementations
- Examples
- Extension

Rationale

- We provide an example as discussed within the GIANT consortium. Consider two studies with sample sizes 32000 and 8000 both with p values 1e-8, we have a combined two-sided p value of 1.49e-14 but also yields p=4.89e-8 with p1=1e-4 and p2=1e-5 (weighted z-score method and more results are available from metap in gap).
- In general, it statistically combines data from multiple studies in the consortium to learn about association (level of significance) and factors related to variations in its magnitude (effect size). We have test of significance = size of effect x size of study, e.g., $\chi_1^2 = r^2 N$ (Kramer & Rosenthal. Comprehensive Clinical Psychology 3-15, Elsevier 1998)

Combining independent tests

- Fisher's method
- One can use truncated p values
- Stouffer's method is based on normal approximation.
- The R implementation is straightforward with sum(-2 * log(pvalues))) and sum(qnorm(1-pvalues)) / sqrt(k)).

$$\chi^2_{2k} = -2\ln P_i \quad i = 1,...,k,$$

$$z = 1/\sqrt{K}\,\Sigma_{i=1}^{k}\Phi^{-1}(1-P_i)$$

- Fisher's method has limitations in
  - Giving equal weight to studies with different sizes
  - No test of heterogeneity
  - No point estimate to become more precise as K increases
- However, there is suggestion about bias regarding msSNP.

Regression models for meta-analysis

- Fixed effects model is unable to account for heterogeneity since deviations from θ and θ are assumed to be explained by random error.
- Random effects model. It is assumed that each study has its own effect distribution against a common distribution.
- The popular DerSimonian-Laird (DL, moment) estimator equates the expectation of the heterogeneity statistic.
- We can include covariates in the model to make study-specific adjustments, i.e., meta-regression.
- Simple heterogeneity (SH) model uses GLS with strictly positive variance estimate.

$$\theta_i = \theta + \varepsilon_i, i = 1,...,k,$$

$$\varepsilon_i \sim N(0, \sigma_i^2)$$

$$\theta_i = \theta + b_i + \varepsilon_i, i = 1,...,k,$$

$$b_i \sim N(0, \tau^2),\ \varepsilon_i \sim N(0, \sigma_i^2)$$

$$\hat{\tau} = \frac{\sum_{i=1}^{k} \sigma_i^{-2}(\theta_i - \hat{\theta})^2 - (k-1)}{\sum_{i=1}^{k} \sigma_i^{-2} - \sum_{i=1}^{k} \sigma_i^{-4} \Big/ \sum_{i=1}^{k} \sigma_i^{-2}}$$

$$\theta_i = \theta + \theta_1 z_i + b_i + \varepsilon_i, i = 1,...,k,$$

$$b_i \sim N(0, \tau^2),\ \varepsilon_i \sim N(0, \sigma_i^2)$$

$$Var(\hat{\theta}_i) = \tau^2 + \sigma^2 = \tau^2(r_i + 1)$$

$$E(\hat{\theta}) = X\theta, Var(\theta) = \tau^2 V$$

Measure of Heterogeneity

- Cochran's Q, $Q = \sum_{i=1}^{k} (\theta_i - \bar{\theta})^2 / \sigma_i^2$, can be referred to a chi-squared distribution with k-1 degrees of freedom.
- I^2, defined as 100%(Q-df)/Q, therefore expresses the percentage of between-study variability that is attributable to heterogeneity rather than chance. Thresholds of 20%, 50%, and 75% are suggested to have low, moderate and high heterogeneity (Higgins *et al*. *BMJ* 2003; 327:57-60).
- It has been suggested that $cQ \sim x^2(v)$ with Q being heterogeneity chi-square, has excellent property (Bohning et al. 2008).

Implementations

- SAS has no built-in procedure for meta-analysis but can customarily done via PROCs GLM (fixed effects/inverse variance) and more often MIXED as well as macros.
- Stata has a comprehensive collection of meta-analysis, notably metan.
- R hosts several package at CRAN (e.g., meta, rmeta) .
- S-PLUS has user-written packages, e.g., hblm.
- Others such as HLM, MLwiN, WinBUGS.
- Customized programs

Useful URLs

- CAMAN (Computer Assisted Analysis of Mixtures)
  http://www.charite.de/biometrie/schlattmann/book/
- improved.ci (function for the improved confidence interval using DL method)
  http://www.statistik.tu-dortmund.de/ma_book.html
- MiMa (An S-Plus/R Function to fit Meta-Analytic Mixed, Random, and Fixed-Effects Models)
  http://www.wvbauer.com/downloads.html
- Hblm (Hierarchical Bayes Linear Model Programs)
  ftp://ftp.research.att.com/dist/bayes-meta/
- CAMAP (Computer-Assisted Meta-Analysis with the Profile Likelihood)
  www.reading.ac.uk/~sns05dab/software.html

Variance of heterogeneity

- The improved confidence interval for the DL estimator (Hartung et al. 2008) is

  improved.ci(logrr, se, 0.05)
- Various estimates of the heterogeneity variance according to Viechtbauer including that for SH (Schlattmann 2009).

  source("mima.ssc")

  covar <- c()

  var <- se^2

  mima(logrr,var,covar,method="SH") # DL

SAS

```
PROC MIXED method = ml data=gwa;
    class study;
    model beta = / s cl;
    repeated / group = study;
    parms / parmsdata=gwa eqcons = 1 to 20;
run;
PROC MIXED data=test covtest;
    class trial;
    model ln_or = / s cl outp=predp outpm=predm;
    repeated diag / R;
    random trial / G gdata = test s v;
    ods output CovParms=cp G=G R=R V=V SolutionF=SF
   SolutionR=SR;
run;
```

Stata

- usesas using meta5.sas7bdat, clear
- reshape long b se, i(locus)
- metan b se, by(locus) fixedi nograph

R

- library(foreign)
- setwd(".")
- meta5 <- read.dta("meta5.dta")
- attach(meta5)

- library(meta)
- by(meta5,locus,function(x) metagen(b,se,data=x))

WinBUGS

```
model
{
  for (i in 1:r)
  {
    y[i] ~ dnorm(psi[i],w[i])
    psi[i] ~ dnorm(theta,t)
  }
  theta ~ dnorm(0,1.0E-4)
  t ~ dgamma(0.001,0.001)
  tausq <- 1/t
}

list(y = c(0.864, 0.646, 0.272, 0.916, 0.867, 0.819, 0.809, 1.212,
     -0.273), w = c(4.40, 9.89, 16.81, 8.38, 8.15, 10.36, 10.79, 4.40,
    15.95), r = 9)
list(theta = 0, t = 1, psi = c(0,0,0,0,0,0,0,0,0))
```

Customized programs

- METAL
- MetABEL
- R/snpMatrix

Examples

```
library(CAMAN)
data(aspirin)
aspirin
mix <- mixalg(obs="logrr", var.lnOR="var", data=aspirin)
library(rmeta)
attach(aspirin)
annotate <- cbind(name,year)
metaplot(logrr,se,labels=annotate)
library(meta)
mg <- metagen(logrr,se)
plot(mg)
funnel(mg)
metabias(mg, method=“linreg”)
```

GWAS consortia

- METAL
- GIANT, BPGen, LFGC

A cautionary note

- In a meta-analysis, we compute effect size for each study and combine them but not combine summary data and compute an effects size for the combined data.
- This allows for a check of consistence regarding effect sizes across studies and minimizes the potential confounders.
- If we were to pool data across studies and then compute the effect size from the pooled data, we may get the wrong answer, due to Simpon's paradox.

See Chapter 13 of Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis.* Wiley 2009

Extensions: multivariate Meta-Analysis

- Background
  - A gene-based association testing (Neale & Sham) not dissimilar to the usual Fisher p value method
  - Multilocus scan statistics (Hoh & Ott) not taking off
  - Bayesian meta-analysis is more involved and the formulation via summary data as in Verzilli *et al*. is not not necessarily used.
  - P values adjusted for correlated tests (p_ACT, Conneely & Boehnke) addresses the following question: What is the minimum p value and more importantly given it is obtained what are the significant levels for all others?
- Problems
  - Covariance of association tests can be poorly estimated given multicollinearity between SNPs at a region/gene.

Statistical models

- The data typically involve *b*, *SE* from linear regression of nearby SNPs to allow for fixed- and random effects modeling and assessment of statistical significance.
- It is not obvious how to infer covariance matrix involving these *b*'s. However, we can work around with respect to pair-wise correlations (*r*).
- For linear regression, it is known that *r* and *t* (=*b*/*SE*) is related via a simple expression $r^2 = t^2/(n-2+t^2)$.
- The covariance between pair-wise correlation has the following form.

Covariance between pairs of correlations

$$Cov(r_{st}, r_{ut}) = [0.5\rho_{st}\rho_{ut}(\rho_{su}^2 + \rho_{st}^2 + \rho_{tu}^2 - 1) + \rho_{su}(1 - \rho_{st}^2 - \rho_{ut}^2)]/n$$

$$Cov(r_1, r_2) = [\rho_1\rho_2\rho_{12}^2 + (\rho_1^2 + \rho_2^2 - 1)(\rho_1\rho_2 - 2\rho_{12})]/2n$$

$$\text{or } (1 - \rho^2)^2/n \text{ with } \rho_1 = \rho_2 \equiv \rho \text{ and } \rho_{12} = 1$$

Combination of SNPs via GLS

- The results of k independent studies, each with p correlations, can be expressed as the concatenation of the vectors of all available correlations. The large sample variance-covariance matrix is then block diagonal. The estimation of the pooled correlation matrix can then be done via weighting or via a generalized least squares (GLS) framework.
- A test of homogeneity of correlation matrices among studies can be performed (Becker 1992). We can accommodate the heterogeneity via a random effects model such that population correction for specific study is a result of the population correlation and study specific factor.
- The implementation (e.g., in R) accounts for variable number of SNPs from each study (Verzilli et al. 2008).

p_ACT and p_ACT_meta

- p_ACT is based on multivariate normal (MVN) assumption originally for sample with individual genotypes but recently extended to results from consortium meta-analysis.
- The basic idea with p_ACT_meta is to find the minimum p value from the collection of correlated SNPs and obtain subsequent p values based on MVN conditional distributions (Holm's procedure) using R/mvtnorm.
- It uses a James-Stein shrinkage estimate as implemented in R/corpcor. A description of mvtnorm appears in *The R Journal*.
- However, the omnibus approach noted earlier is appealing.

References

- Bohning D, Kuhnert R, Rattanasiri S. *Meta-Analysis of Binary Data Using Profile Likelihood*. CRC Press 2008
- Conneely KN, Boehnke M. *AJHG* 2007;81:1158-68
- Demidenko E. *Mixed Models*. Wiley 2004
- Harris *et al. Stata J* 2008; 8:3-28
- Hartung J, Guido K, Sinha BK. *Statistical Meta-Analysis with Applications*. Wiley 2008
- Normand S-L. T. Stat Med 1999;18:321-59
- Rao DC, Gu CC. *Genetic Dissection of Complex Traits*, 2nd Ed. Academic Press 2008
- Schlattmann P. *Medical Applications for Finite Mixture Models*. Wiley 2009
- Sidik & Jonkman. *Appl Stat* 2005; 54:367-84
- Sterne J. Meta-Analysis in Stata. Stata Press 2009.
- Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Wiley 2000
- Verzilli *et al. AJHG* 2008; 82:859-72
- Whitehead A. *Meta-Analysis of Controlled Clinical Trials*. Wiley 2002

Summary

- It is not intended to provide a comprehensive overview but simply offer some flavour of the kind of thinking and practice.
- Evidence synthesis with conscious recognition of heterogeneity is in the heart of meta-analysis.
- Fixed effects analysis is restricted to data of the type found in the studies included, but random effects model generalizes to all studies of the type from which our studies were drawn. Results from both models together with SH model are highly recommended.
- We have omitted the graphical aspects, e.g., Bax *et al. AJE* 2009; 169:249-55. An Excel macro is available from http://www.mix-for-meta-analysis.info/index.html

IMS Institute of Metabolic Science

MRC | Epidemiology Unit

Genome-wide Association Studies

Other topics

Jing Hua Zhao

Topics

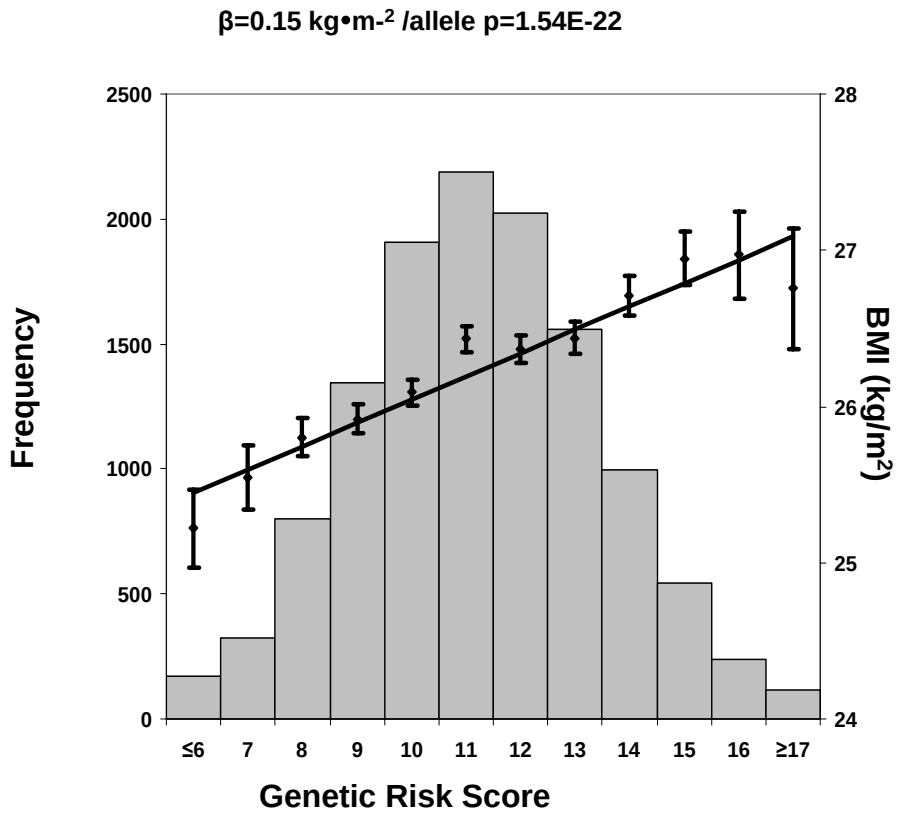
- Risk prediction
- Instrumental variable method
- Structural equation modeling
- Building networks
- Analysis of family data

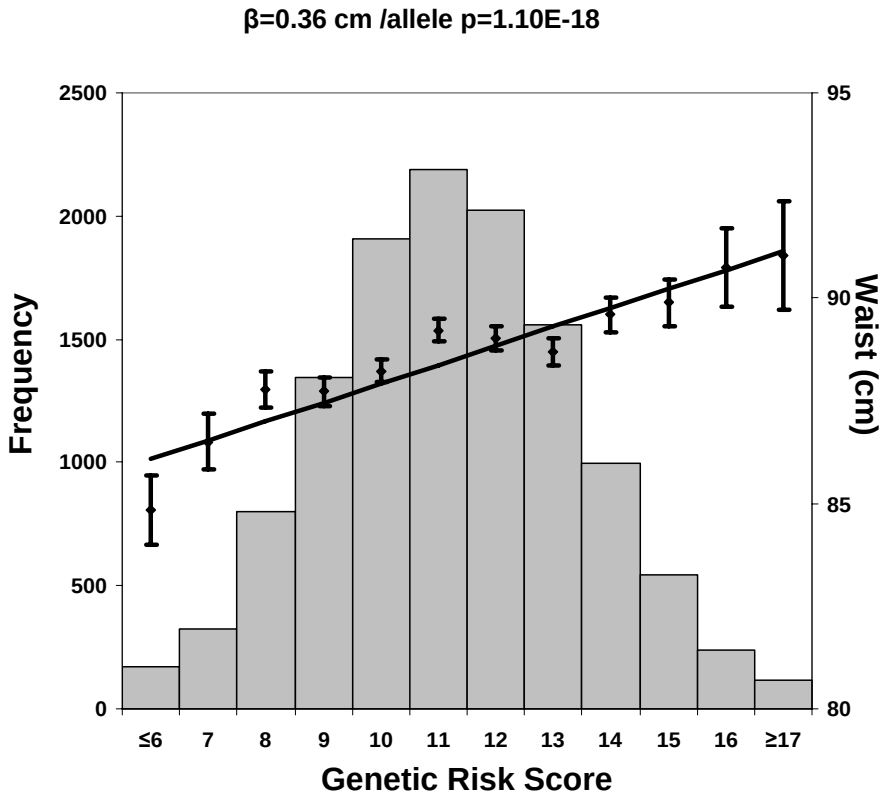
Risk prediction

- A set of SNPs can be used in a logistic regression model to predict if an individual is a case or control based on a cut-off probability. An optimal cut-off can be facilitated through receiver operating characteristics (ROC) curve. The ability to classify individuals correctly is measured by area under the ROC curve (AUC, e.g. ~0.5, 0.7-0.8, 0.8-1 for no, acceptable, excellent discrimination).
- Examples: prostate cancer, obesity, HDL/TG/LDL.
- A testing example

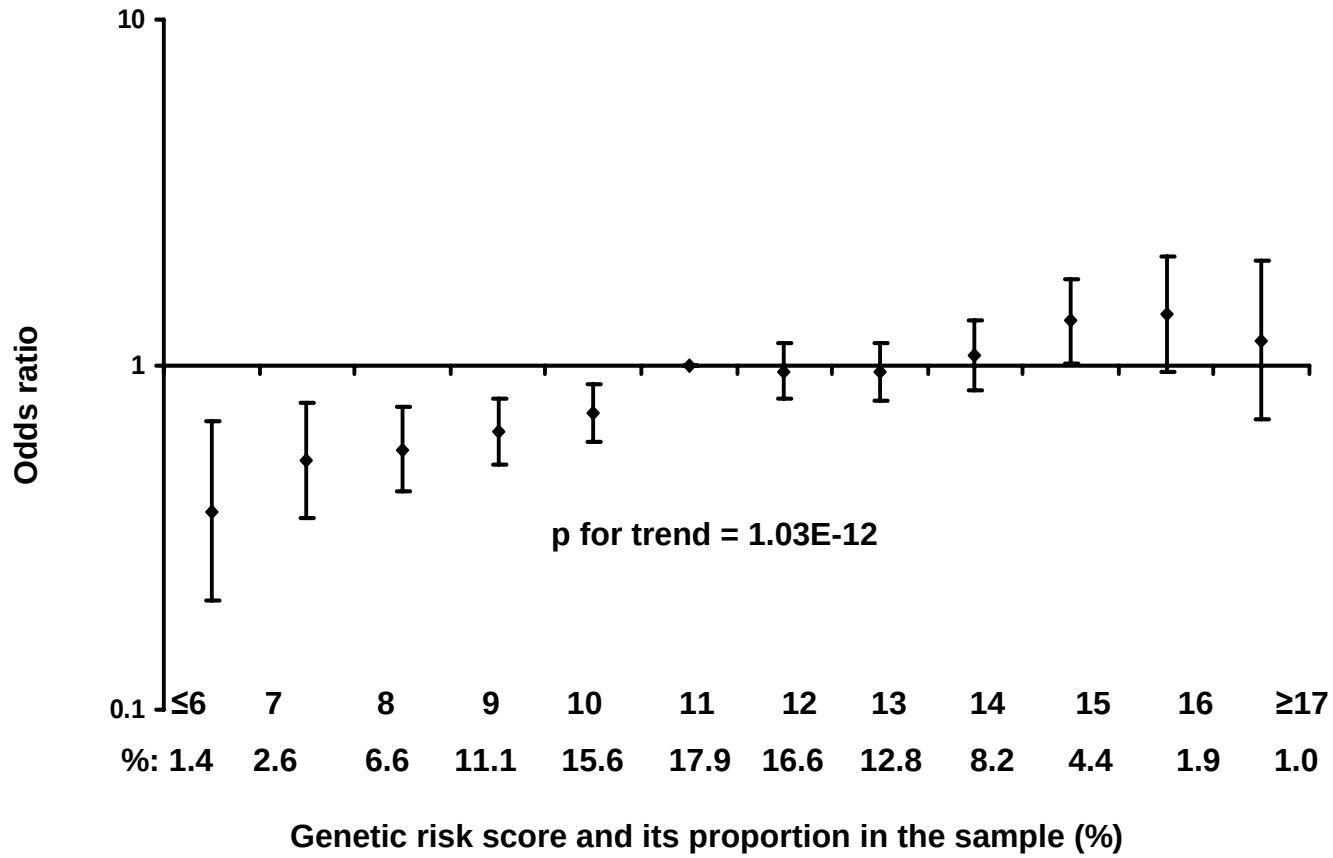
```
library(verification)
obs<- round(runif(100))
pred<- runif(100)
A<- verify(obs, pred, frcst.type = "prob", obs.type = "binary")
roc.plot(A, main = "Test", binormal = TRUE, plot = "both")
roc.plot(A, threshold=seq(0.1,0.9, 0.1), CI=TRUE, alpha=0.1)
roc.plot(obs,pred,xlab='1-specificity',ylab='sensitivity',cex=2)
AUC <- roc.area(obs,pred)$A
```

Risk score and BMI in EPIC-Norfolk

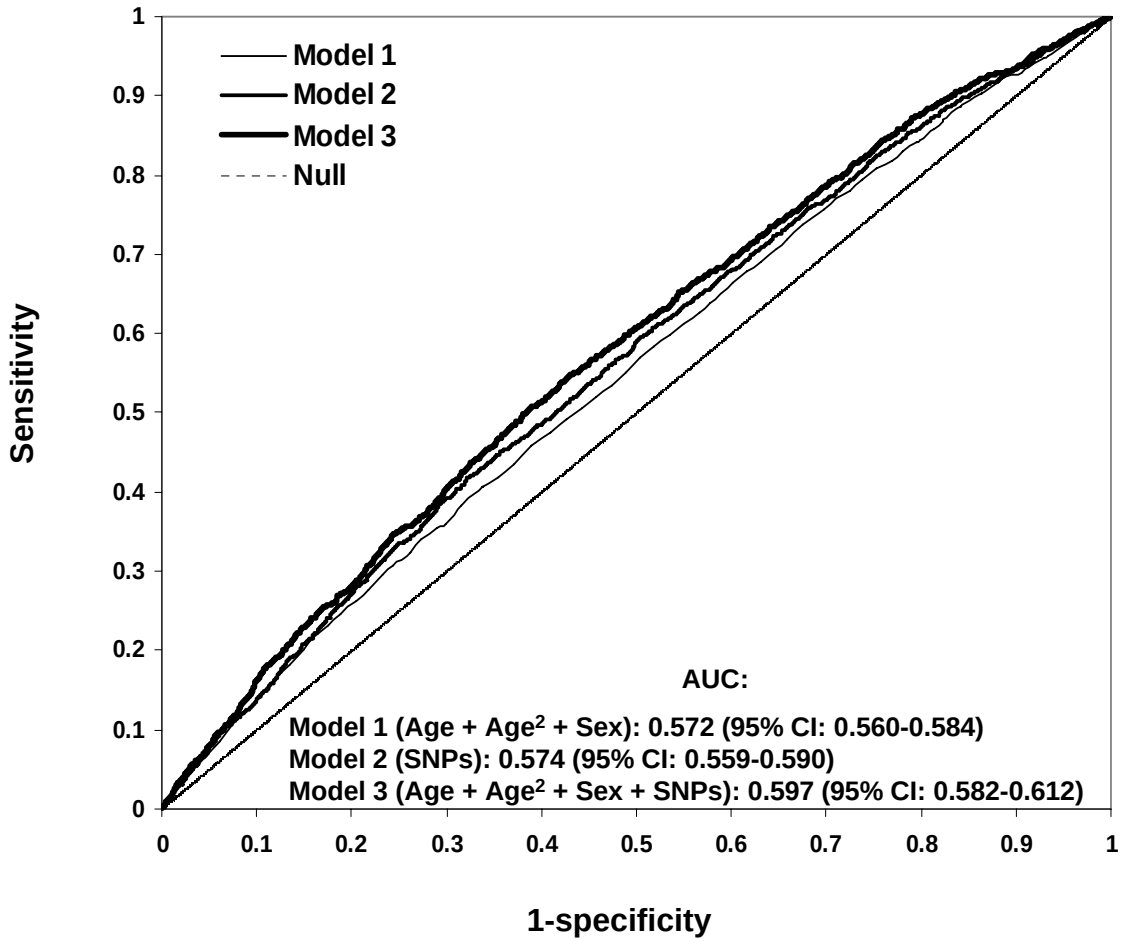




Risk score and obesity/overweight



ROC curve and AUC

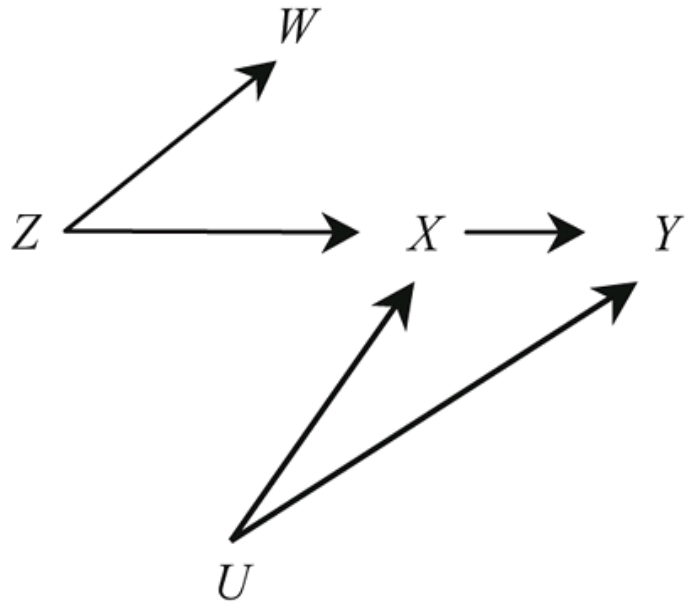


Instrumental variable (IV) estimation

- It is a method for estimating regression $Y=(Z\ X)b+e$ parameters b when X are measured with error, $W=X+U$, and possibly when a second or biased but independent measurement (T) is available. Given $cov(T,e) = cov(T,U)=0$, $cov(T,X)<>0$, $b=cov(T,Y)/cov(T,W)$.
- More formally, 1. T is uncorrelated with X; 2. T is independent of the measurement error $U=W-X$ in the surrogate W; 3. (W,T) is a surrogate for X so that $E(Y|Z,X,W,T)=E(Y|Z,X)$.
- See Fuller WA. Measurement Error Models. Wiley 1987; Greene WH. Econometric Analysis, 5th Ed. Prentice Hall 2003; Carroll et al. Measurement Error in Nonlinear Models-A Modern Perspective, 2nd Ed. CRC 2006; Gelman A, J Hill. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press 2007

IV in simple terms

In an observational study, U represents unmeasured confounders of the X–Y association. In a randomized trial, U represents variables that affect adherence to treatment assignment and thus influence received treatment X. Z is called an instrumental variable (or instrument) for estimating the effect of X on Y.

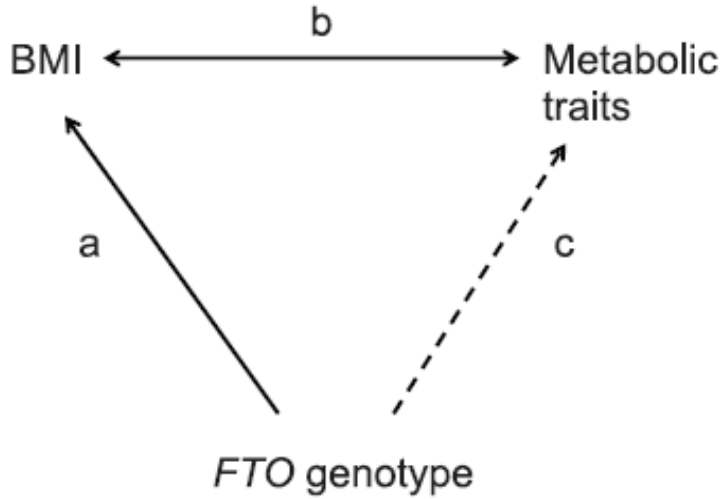


Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd Edition, Lippincott Williams & Wilkins 2008

a. Z affects X (i.e., Z is an ancestor of X). b. Z affects the outcome Y only through X (i.e., all directed paths from Z to Y pass through X). c. Z and Y share no common causes.

FTO genotype, BMI and T2D

- There is epidemiological association between BMI and metabolic traits.
- There is association between *FTO* and BMI.
- The association between *FTO* genotype and metabolic traits would be mediated by BMI (c=axb).



- This is the so-called triangulation approach (Freathy et al. Diabetes 2007; 57:1419-26).

Two-stage least squares

- We can lay out two equations

$$TG = \beta_0 + \beta_1 BMI + \beta_2 SNP + error$$
$$BMI = \gamma_0 + \gamma_1 SNP + error$$

- We can plug in the second equation into the first.

$$TG = \beta_0 + \beta_1 BMI + \beta_2 SNP + error$$
$$= \beta_0 + \beta_1(\gamma_0 + \gamma_1 SNP) + error$$
$$= (\beta_0 + \beta_1\gamma_0) + (\beta_1\gamma_1 + \beta_2)SNP + error$$

- We proceed with two steps:

1. We first regress TG on SNP.
2. We also regress BMI on SNP.

$$TG = \delta_0 + \delta_2 SNP + error$$
$$BMI = \gamma_0 + \gamma_2 SNP + error$$

- We then have the Wald estimate with $\beta_2 = 0$

$$\gamma_2 = \beta_1\gamma_1 + \beta_2$$
$$\beta_1 = (\delta_1 - \beta_2)/\gamma_1$$

- A summary in our setting is Bochud et al. *IJE* 2008, 37:414-6

Structural equation modeling

- Several examples seen in recent GWAS literature can be modeled via path analysis or put in this framework.
- It is typically a confirmatory analysis based on model-fitting.
- It has been a rather useful tool to study causal relationship.
- It is natural to study change using longitudinal data.
- sem package in R is a very good initiative, but it is often necessary to resort to other systems such as EQS, AMOS, M*plus,* e.g., the inter-relationship between anthropometric measurements using M*plus*.
- A critique is that SEM relies on conditional independence assumptions with IV being as a special case, so that the assumptions required for causal effects are difficult to satisfy. It is helpful to examine equivalent models.

M*plus* code

```
Title:
 snp1: rs1121980 from FTO
 snp2: rs17782313 from MC4R
 zlbmi : BMI
 zlwst : waist
 zltg : Triglycerides
 zsys : SBP
 zdia : DBP
Data:
 File is effectsize.dat ;
Variable:
 Names are snp1 snp2 zlbmi
zlwst zltg zsys zdia;
 Missing are all (-9999) ;
 Usevariables are snp1 zlbmi zltg;
```

```
Model:
 zltg on zlbmi;
 zlbmi on snp1;
 zltg on snp1;
Model indirect:
 zltg ind snp1;
Output:
 Standardized;
```

Bayesian networks

- Rule-based systems with certainty factors have serious limitations as a method for knowledge representation and reasoning under uncertainty, and attention towards a probabilistic interpretation of certainty factors leads to Bayesian networks.
- It can be described briefly as an acyclic directed graph (DAG) which defines a factorization of a joint probability distribution over the variables represented by the nodes of the DAG.
- The process of construction involves identification of the relevant variables and their causal relations, which leads to DAG specified in terms of a set of conditional probabilities.
- Example: Bayesian networks for gene expression data in GAW15 problem 1.

References

- Krzanowski WJ, Hand DJ. ROC Curves for Continuous Data. CRC 2009
- Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press 2003
- Gonen M. *Analyzing Receiver Operating Characteristic Curves with SAS*, SAS Institute Inc. 2007
- Loehlin JC. *Latent Variable Models-An Introduction to Factor, Path, and Structural Equation Analysis*. 4th Edition, Lawrence Erlbaum Associates 2004
- Kline RB. *Principles and Practice of Structural Equation Modeling*. 2nd Edition, The Guiford Press 2005
- Gentleman R, et al. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer 2005.
- Bollen KA, PJ Curran. *Latent Curve Models-A Structural Equation Perspective*. Wiley 2006
- Kjaerulff UB, AL Madsen. *Bayesian Networks and Influence Diagrams-A Guide to Construction and Analysis*. Springer 2008
- Emmert-Streib F, Matthias D. *Analysis of Microarray Data-A Network-Based Approach*. Wiley-VCH 2008
- Junker BH, F Schreiber (Ed). *Analysis of Biological Networks*. Wiley 2008

Example: GAW16 Framingham data

- Data management through SAS
- QC and basic association statistics via PLINK
- Estimation of inflation factor by snpMatrix
- Cross-check with GRAMMAR procedure from R/GenABEL
- Longitudinal data with SAS, Stata and Mplus
- Graphics via R/gap

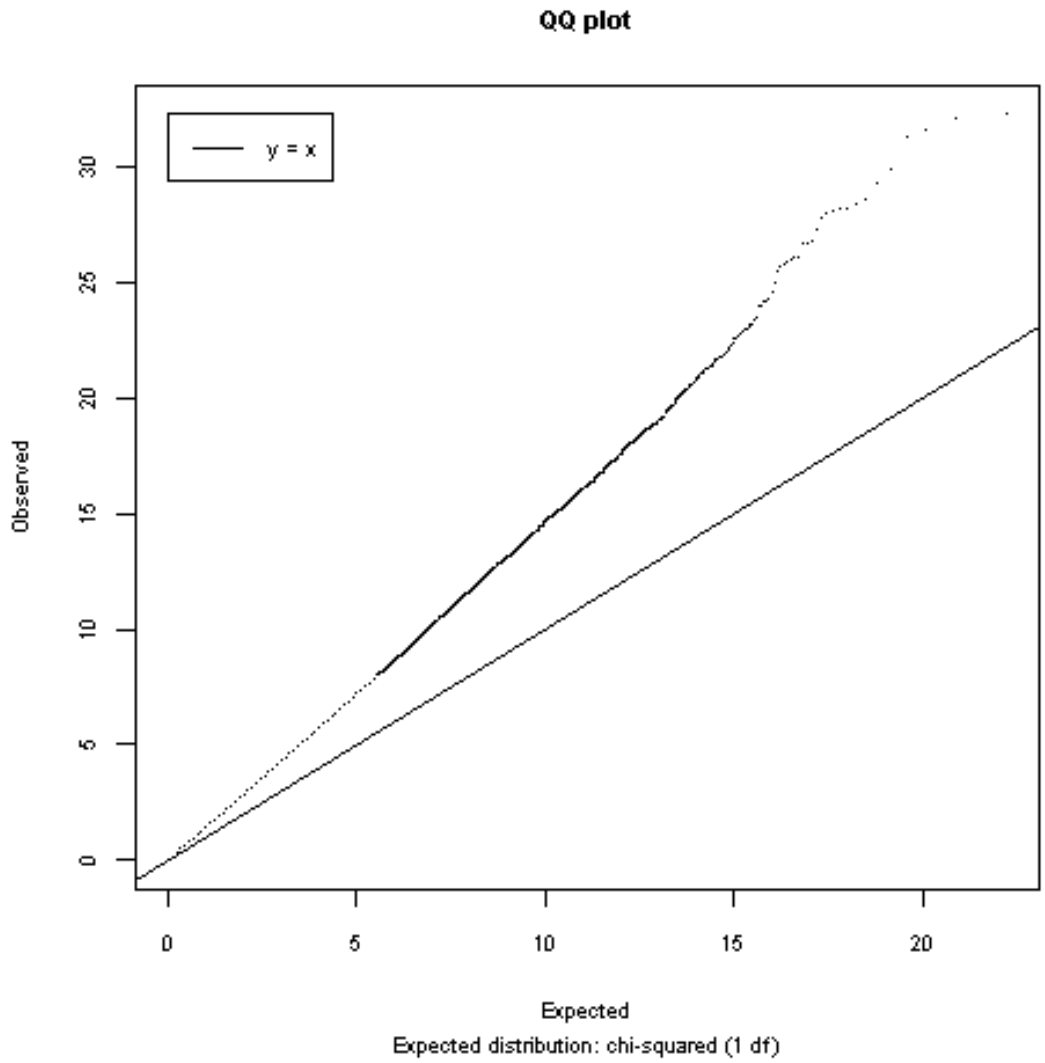
GenABEL: data input

```
library(GenABEL)
# this is an example of Framingham data for GAW16
convert.snp.tped(tped = "chrall.tped", tfam =
    "pheno.tfam", out = "chrall.raw", strand = "+")
df <- load.gwaa.data(phe = "pheno.dat", gen =
    "chrall.raw", force = TRUE)
```

IBS and polygenic model

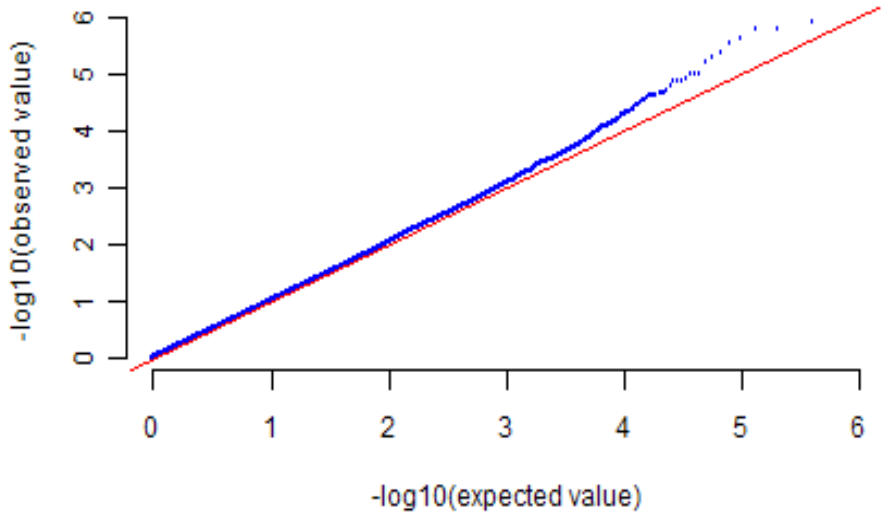
```
gkin <- ibs(df@gtdata,w="freq")
save(gkin, file="kin.Rdata")
pg <- polygenic(bmi1 ~ sex + age, kin=gkin, df,
   quiet=TRUE)
pgres <- pg$res
write(pgres, file="genabel.dat", 1)
save(pg, file="bmi.Rdata")
```

Q-Q plot of the original p values

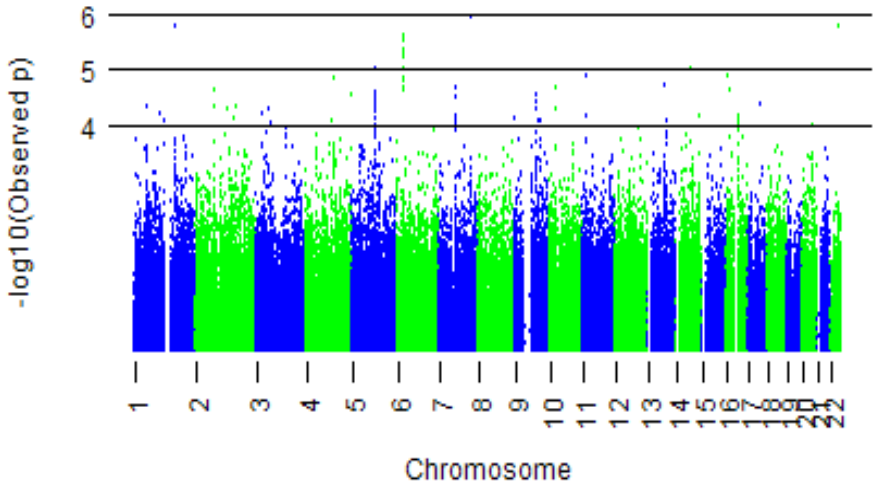


Results with genomic control (λ=1.425)

Q-Q plot



Manhattan plot



Further example – expression quantitative trait

- There is substantial individual variation in expression level of genes, which is smaller in monozygotic twins than among individuals of other relationships, suggesting a genetic component.
- Genetic Analysis Workshop 15 problem 1 provided
  - 14 three-generation families
  - 2554 expression quantitative traits
  - 2882 SNP genotypes
  - Chromosomal positions of these SNPs
- These information was contained in comma-delimited files each with appropriate header. This simple example serves to illustrate the basic analysis involved.

Getting data into R

- We used the following code,

```
id <-
    read.table("LINKAGE.PED",header=TRUE,as.is=TRUE,sep=",")
phn <-
    read.table("LINKAGE.PHN",header=TRUE,as.is=TRUE,sep=",")
snp <-
    read.table("LINKAGE.SNP",header=TRUE,as.is=TRUE,sep=",",
    na.string="0/0")
map <-
    read.table("LINKAGE.MAP",header=TRUE,as.is=TRUE,sep=",")

pheno <- merge(id,phn,by=c("FAMID","ID"))
ped <- merge(pheno,snp,by=c("FAMID","ID"))
```

- Now the object ped has all the information as required

Summary statistics

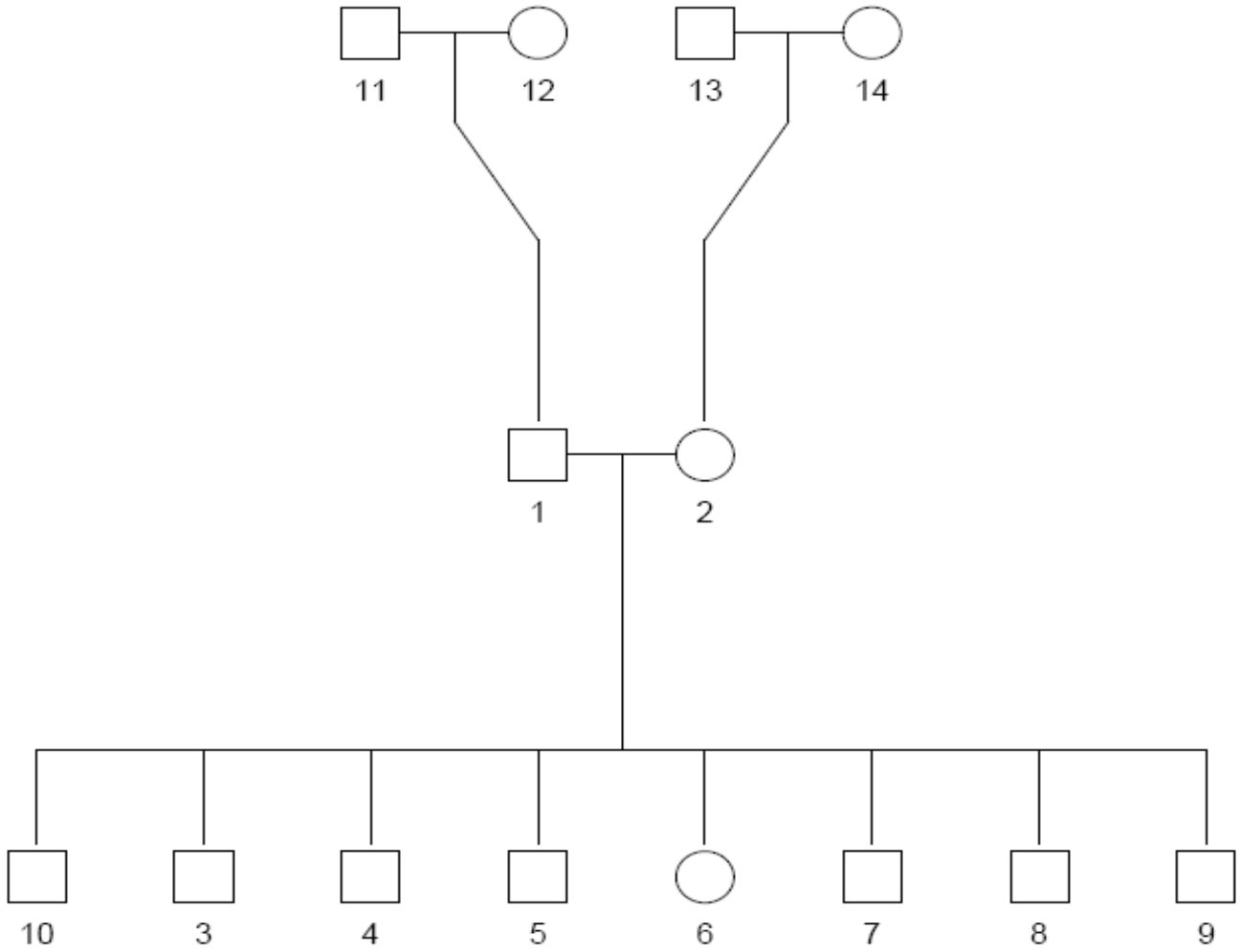
```
dim(pheno)
>194 3559
dim(snp)
>194 2884
dim(ped)
>194 6441
with(pheno,table(FAMID))
FAMID
1333 1340 1341 1345 1346 1347 1362 1408 1416 1418 1421
   1423 1424 1454
  14   13   14   13   14   14   14   14   14   14   14   14   14   14
```

- There were 194 individuals in 14 families, for which 2882 SNPs were available.

Pedigree diagrams

```
library(kinship)
pdf("pedfile.pdf"); attach(ped)
uid <- unique(ped$FAMID)
for (j in 1:length(uid))
{
  selected <- FAMID==uid[j]
  id <- ID[selected]
  dadid <- FA[selected]
  momid <- MO[selected]
  sex <- SEX[selected]
  par(xpd=TRUE)
  ped <- pedigree(id, dadid, momid, sex)
  plot(ped, id=paste("\n",id,sep=""))
  title(uid[j])
  k <- kinship(id,dadid,momid)
  print(k)
}
detach(ped); dev.off()
```

A typical pedigree diagram



Analysis with SNPassoc

```
> library(SNPassoc)
> SNPs <- setupSNP(ped,3560:6441,sep="/",info=map)
> founders <- subset(SNPs,FA+MO==0)
> hwe <- tableHWE(founders)
> chr22 <- subset(map,CHROMOSOME==22)[,2]
> ld22 <- LD(founders,chr22)
```

- Note that the argument info for setupSNP() is optional, and that founders are individuals who have no data for parents available. It is appropriate to perform Hardy-Weinberg equilibrium test and marker-marker association only in founders. We use 57 SNPs in chromosome 22 for illustration.

Summary statistics

```
> summary(ld22)
     Length Class  Mode
call       3   -none- call
D       3249   -none- numeric
D'      3249   -none- numeric
r       3249   -none- numeric
R^2     3249   -none- numeric
n       3249   -none- numeric
X^2     3249   -none- numeric
P-value 3249   -none- numeric
> ld22$"R^2"
```

- It can be seen that there are 57*57=3249 items in seven summary statistics.

Single-SNP association

```
> association(X1007_s_at ~ SEX+rs738842,model="log-
  additive",data=SNPs)

SNP: rs738842  adjusted by: SEX
                  dif    lower  upper p-value   AIC
log-Additive
0,1,2          0.09292 -0.02632 0.2122  0.1267 289.7
```

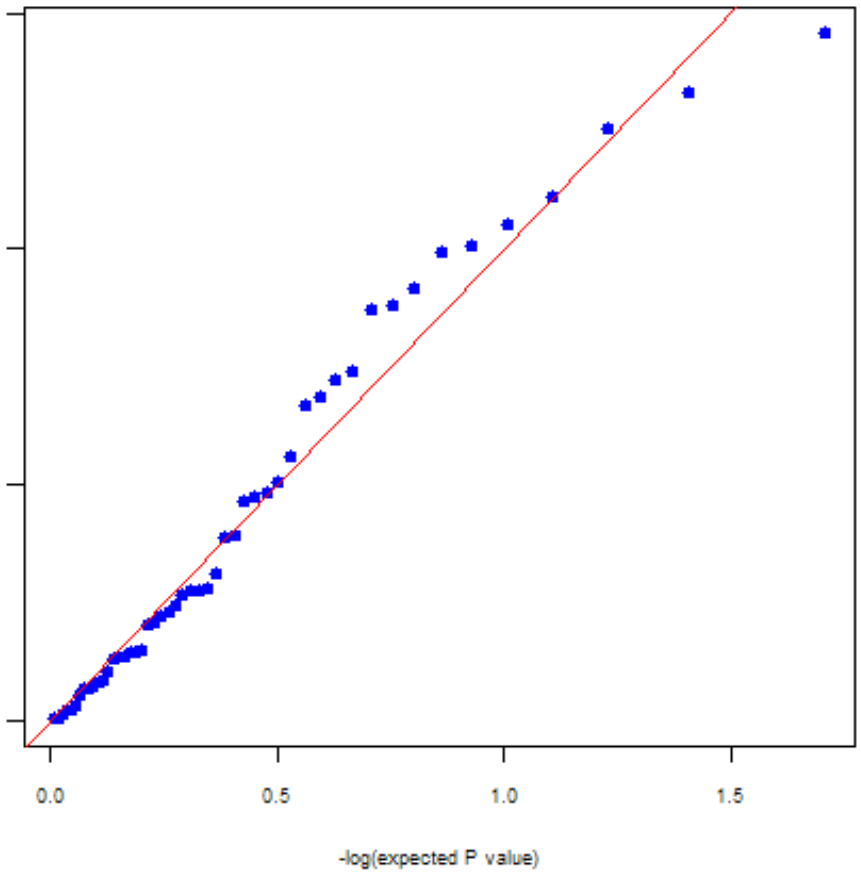
- We give SNP rs738842 for illustration. It is a regression model with SEX as a covariate in the model.

All SNPs in chromosome 22

```
> wg22 <- WGassociation(X1007_s_at ~
  SEX+1,model="log-
  add",data=founders[,c(5,6,6385:6441)])
> summary(wg22)
> png("chr22.png")
> qqpval(wg22$"log-additive")
> dev.off()
```

- We still use chromosome 22 SNPs only. We use the qqpval() function for Q-Q plot.

Q-Q plot



Formal use of family data

- In the context of GWAS, it is appropriate to use general statistical framework such as models for longitudinal data.
- In R, this can be done using packages such as gee, nlme.
- Therefore we omit the details and leave it as an exercise.

Some reflections

- GWAS is more useful than candidate gene alone by providing prior information of test statistics as shown with the Framingham data.
- GWAS goes beyond candidate gene studies in that it allows for analyses of a variety of traits.
- Our use of case-cohort design allows for study of events, not as population-based controls.

Acknowledgement

- Shengxu Li for the EPIC-Norfolk risk score plots
- Colleagues at MRC for support and help
- GAW15 for the CEPH data
- GAW16 for Framingham data

Institute of Metabolic Science
MRC Epidemiology Unit

Genome-wide Association Studies

Summary

Jing Hua Zhao

General comments

- The R environment for genetic association analysis can be an integral part of analysis for GWAS.
- It has achieved a great deal of stability though somewhat evolving, as is indicated by the fact that we still need to tune our data into the required format for individual packages, but this has been greatly improved. We have used a lot of BASH or scripts coupled with R to do so. The R environment and other software should be complementary with each other.
- Graphics has facilitated our understanding tremendously, but is still in need of dynamic elements in the presentation.
- A range of models available from R remains to be explored and at least they can be supplementary to the main analysis for GWAS.

Scientific issues of GWAS

- What are the uses of GWASs?
  - Discovery of new susceptibility loci
  - Elucidate biologic pathways
  - Identify links between these loci and covariates
  - Risk prediction
- Where would GWASs go?
  - Expanding well-characterised study populations
  - Expanding the range of genetic variation including structural variants and lower-frequency common variants
  - Documenting functional mechanisms responsible for the association signals.

Chanock (*Personal Communication*) and Altshuler et al. *Science* 2008, 322:881-8

Limitations and practical issues in GWAS

- Limitations
  - It requires large sample sizes
  - It only identifies loci, not genes
  - It detects only common alleles in a population
  - It usually does not go into the expression level
- Practical issues
  - For individual GWAS, the issues as of epidemiological studies in general remain and there is uncertainty to declare statistical significance
  - For consortium meta-analysis, there may be difference in quality control, data sharing and variation in complexity of analysis

References

- R-oriented
  - Foulkes AS. *Applied Statistical Genetics with R for Population-based Association Studies*. Springer 2009
  - Gentleman R, V Carey, W Huber, R Irizarry, S Dudoit. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Springer 2005
  - Siegmund D, B Yakir. *The Statistics of Gene Mapping*. Springer 2007
  - Wu R, C-X Ma, G Casella. *Statistical Genetics of Quantitative Traits-Linkage, Maps and QTL*. Springer 2007
- General
  - Lange K. *Applied Probability*. Springer 2003
  - Sorensen D, D Gianola. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer 2002
  - Thomas DC. *Statistical Methods in Genetic Epidemiology*. Oxford 2004

Summary

- The need from analyses in GWAS seeds the development in R and shares much in common with many other problems involving large data, such as interactive graphics in combination with publicly available databases, the use of statistical and computational facilities available from the R system.
- Applications in substantial areas are the constant source of motivation in package development. The implementation is likely to be patchy but with a great prospect, e.g., advanced models and causal pathways.
- We certainly do not wish to restrict ourselves to the R environment alone.